

# A registration method for improving quantitative assessment in probabilistic diffusion tractography

J.L. Waugh<sup>a,c,e,h,j,\*</sup>, J.K. Kuster<sup>a,d,j</sup>, M.L. Makhoulf<sup>a,d,i,j</sup>, J.M. Levenstein<sup>a,d,j</sup>,  
T.J. Multhaupt-Buell<sup>c</sup>, S.K. Warfield<sup>f,h</sup>, N. Sharma<sup>c,g,h,1</sup>, A.J. Blood<sup>a,b,d,h,j,1</sup>

<sup>a</sup> Mood and Motor Control Laboratory, Massachusetts General Hospital, Charlestown, MA, United States

<sup>b</sup> Laboratory of Neuroimaging and Genetics, Massachusetts General Hospital, Charlestown, MA, United States

<sup>c</sup> Dept. of Neurology, Massachusetts General Hospital, Boston, MA, United States

<sup>d</sup> Dept. Psychiatry, Massachusetts General Hospital, Boston, MA, United States

<sup>e</sup> Division of Child Neurology, Boston Children's Hospital, United States

<sup>f</sup> Department of Radiology, Boston Children's Hospital, United States

<sup>g</sup> Department of Neurology, Brigham and Women's Hospital, Boston, MA, United States

<sup>h</sup> Harvard Medical School, Boston, MA, United States

<sup>i</sup> Harvard-MIT HST Program, United States

<sup>j</sup> Martinos Center for Biomedical Imaging, MGH, Charlestown, MA, United States

## ARTICLE INFO

### Keywords:

Tractography  
Registration  
Dice similarity coefficient  
Center of gravity  
Longitudinal  
Quantitative

## ABSTRACT

Diffusion MRI-based probabilistic tractography is a powerful tool for non-invasively investigating normal brain architecture and alterations in structural connectivity associated with disease states. Both voxelwise and region-of-interest methods of analysis are capable of integrating population differences in tract amplitude (streamline count or density), given proper alignment of the tracts of interest. However, quantification of tract differences (between groups, or longitudinally within individuals) has been hampered by two related features of white matter. First, it is unknown to what extent healthy individuals differ in the precise location of white matter tracts, and to what extent experimental factors influence perceived tract location. Second, white matter lacks the gross neuroanatomical features (e.g., gyri, histological subtyping) that make parcellation of grey matter plausible – determining where tracts “should” lie within larger white matter structures is difficult. Accurately quantifying tractographic connectivity between individuals is thus inherently linked to the difficulty of identifying and aligning precise tract location. Tractography is often utilized to study neurological diseases in which the precise structural and connectivity abnormalities are unknown, underscoring the importance of accounting for individual differences in tract location when evaluating the strength of structural connectivity.

We set out to quantify spatial variance in tracts aligned through a standard, whole-brain registration method, and to assess the impact of location mismatch on groupwise assessments of tract amplitude. We then developed a method for tract alignment that enhances the existing standard whole brain registration, and then tested whether this method improved the reliability of groupwise contrasts. Specifically, we conducted seed-based probabilistic diffusion tractography from primary motor, supplementary motor, and visual cortices, projecting through the corpus callosum. Streamline counts decreased rapidly with movement from the tract center (–35% per millimeter); tract misalignment of a few millimeters caused substantial compromise of amplitude comparisons. Alignment of tracts “peak-to-peak” is essential for accurate amplitude comparisons. However, for all transcallosal tracts registered through the whole-brain method, the mean separation distance between an individual subject's tract and the average tract (3.2 mm) precluded accurate comparison: at this separation, tract amplitudes were reduced by 74% from peak value. In contrast, alignment of subcortical tracts (thalamo-putaminal, pallido-rubral) was substantially better than alignment for cortical tracts; whole-brain registration was sufficient for these subcortical tracts.

\* Corresponding author. 120 2nd Ave., Charlestown, MA, 02129, United States.

E-mail addresses: [jeff.waugh@childrens.harvard.edu](mailto:jeff.waugh@childrens.harvard.edu) (J.L. Waugh), [jaaake@gmail.com](mailto:jaaake@gmail.com) (J.K. Kuster), [Miriammakhlouf@yahoo.com](mailto:Miriammakhlouf@yahoo.com) (M.L. Makhoulf), [jacobml@nmr.mgh.harvard.edu](mailto:jacobml@nmr.mgh.harvard.edu) (J.M. Levenstein), [tmulthaupt-buell@mgh.harvard.edu](mailto:tmulthaupt-buell@mgh.harvard.edu) (T.J. Multhaupt-Buell), [simon.warfield@childrens.harvard.edu](mailto:simon.warfield@childrens.harvard.edu) (S.K. Warfield), [nsharma@partners.org](mailto:nsharma@partners.org) (N. Sharma), [ablood@nmr.mgh.harvard.edu](mailto:ablood@nmr.mgh.harvard.edu) (A.J. Blood).

<sup>1</sup> These authors should be considered joint senior authors.



systematic measurement errors if the groups compared differ in tract architecture, yielding non-uniform spatial sampling. For example, if the typical length of a particular tract in a patient population is longer than in healthy controls, the N points of tract measurement will be more widely dispersed in patients and will not sample anatomical sites identical to those in healthy controls. Similarly, if the tract bundle in patients is shifted within the larger white matter skeleton, interval measurements will sample a different local environment (e.g., distinct crossing or parallel tracts) in patients and controls. To be clear, voxelwise comparisons following imperfect registration suffer from the same criticisms, and in fact, along-tract measures are likely more accurate than voxelwise measurements based on whole-brain registration. A theoretical registration method that yielded perfect alignment would outperform both approaches; the point at which a registration is good enough that voxelwise comparisons outperform along-tract measures is unknown.

The explicit goal of registering images into a standard space is to enable quantitative comparison of imaging data across individuals or sessions on a voxelwise basis (e.g., using group averages or programs such as FSL's *randomise*, Winkler et al., 2014). With voxelwise comparison, registration mismatches of even a few millimeters may place the center of tracts in non-overlapping positions. As noted by Jones and Cercignani (2010), intersubject comparisons of diffusion MRI metrics can be substantially compromised by millimeter-scale mismatches in location – which can be introduced at each step of image acquisition, processing, and data extraction. This raises the question of whether quantitative assessments of tractography are truly comparing amplitude at a given location within a tract, or in actuality represent a combination of amplitude and tract location.

In the current study, we describe the variability in tract location with standard non-linear whole-brain registration, and introduce a method for correcting this spatial variance. We illustrate the impact of variability in tract location on quantitative tractographic analyses and demonstrate that registration of tractography to a common “tract space” reduces errors in quantitative voxelwise analysis.

## 2. Materials and Methods

### 2.1. Subjects

Thirty-seven healthy controls with no prior neurologic or psychiatric history were recruited through public advertisement. Data from all 37 subjects were utilized for assessment of tract mismatch (see section 2.6). In addition, fourteen out of the 37 subjects in the study were scanned twice, with one month between scans. This subgroup allowed us to estimate the within-subject test-retest variability in tract location. Twelve subjects served as healthy controls in our prior disease-focused diffusion tractography study (Blood et al., 2012), but have never been assessed via the methods described here. Data from 32 of the same 37 subjects (16 pairs) were utilized to construct artificial, matched cohorts (see section 2.7.2) for simulation of real-world comparisons in groupwise tractography under two conditions: where group differences in amplitude were expected to be present, and expected to be absent. Note that these final groups (which were contrived, constructed by distorting the Gaussian distribution of tract amplitude) were not designed to identify biological reality; rather, we used these groups only to test how location mismatch impacts the accuracy of amplitude measurement.

The mean age of this group was 40.9 years (range: 18–74). Thirty-four subjects were right handed, two were left handed, and one was ambidextrous. Twenty subjects were female; seventeen were male. All participants were provided with printed materials describing the research protocol and were encouraged to ask questions regarding the study. We obtained written consent for all participants. The human studies section of the Institutional Review Board for Partners HealthCare System approved this protocol, including our consent procedure. All research was conducted in accordance with the principles in the Declaration of Helsinki.

### 2.2. MRI acquisition

Subjects were scanned at the MGH Athinoula A. Martinos Center for Biomedical Imaging (Charlestown, MA). Thirty-one healthy controls were scanned on a 3.0 T S Tim Trio MRI (Siemens AG, Medical Solutions, Erlangen, Germany); six subjects were scanned on a Siemens 3.0 T Allegra MRI, as previously described (Blood et al., 2006). In subsequent group analyses, cohorts were matched 1-to-1 for scanner-type, age, gender, and handedness, ensuring that this change in magnet had no impact on group comparisons. For the 14 subjects with repeat scans, the magnet utilized for their first scan was always utilized for their repeat scan. Images were acquired using a high-resolution whole brain DTI sequence with the following sequence parameters: repetition time (TR) = 8 s; echo time (TE) = 83 ms; slice thickness = 2 mm isotropic, 60 slices total, acquisition matrix 1,286,128 [2,566,256 mm field of view (FOV)], six averages, 60 noncolinear directions, with b-value = 700 s/mm<sup>2</sup>, and one image with b-value = 0 s/mm<sup>2</sup>. DTI scans in each subject were acquired using auto-align software (van der Kouwe et al., 2005) to normalize brain image slice orientation between participants and across scan sessions.

### 2.3. Image processing

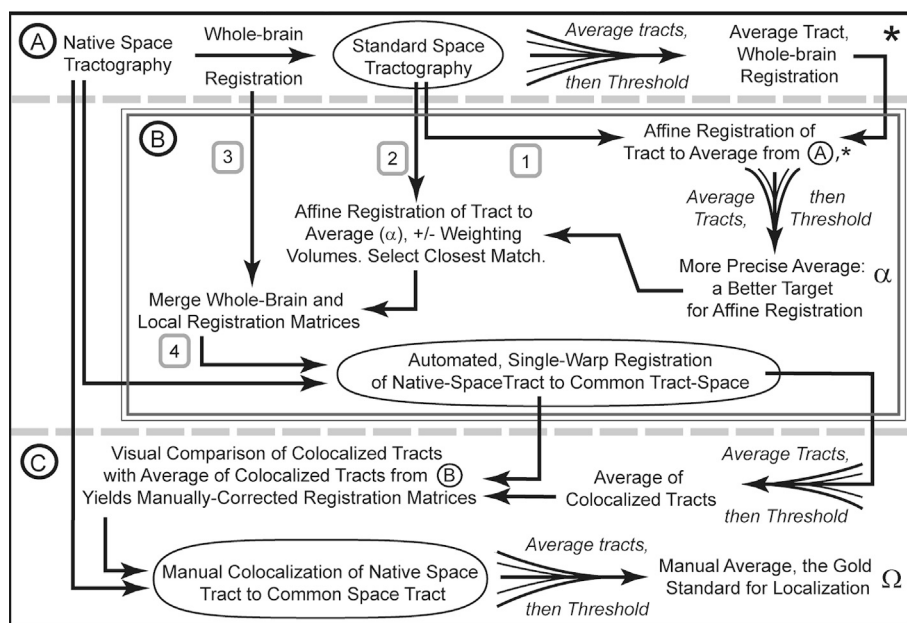
All image analysis was performed using tools from The Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB) software ([www.fmrib.ox.ac.uk/fsl](http://www.fmrib.ox.ac.uk/fsl)) version 5.0, using standard parameters, as previously described (Blood et al., 2012). Specific FSL commands are listed in italics.

**Preprocessing.** Initial data preprocessing for each subject included radiological orientation (*fsorient*), correction for head motion and eddy current distortions (*eddy\_correct*), and removal of non-brain tissue (*bet*). Motion correction parameters from *eddy\_correct* were used in subsequent steps to assess for the impact of motion on variability in tract location. Local diffusion tensors were fit using *dtifit*, creating a 3D FA image at the same matrix size and resolution as the original diffusion images. To generate diffusion parameters at each voxel, *bedpostx* (Behrens et al., 2007) was run on preprocessed DTI data in each subject's native space.

**Standard registration.** FA maps were registered into standard space using serial affine (*flirt*, Jenkinson et al., 2002) and non-linear (*fnirt*) transforms, registering to the FMRIB58\_FA\_1 mm template (Fig. 1A). All whole-brain registrations were visually inspected; all were of high quality and did not require further correction. The resulting whole-brain registration matrix was used in subsequent steps to register subjects' tractography from native into standard space. The inverse of this registration matrix (generated by *invwarp*) was used to register standard space regions of interest (ROI) segmentations into native space for subsequent tractography. Note that this registration was a baseline step applied to all native-space FA images and tracts, and was the first step for refinement in registration detailed in section 2.5.

### 2.4. Probabilistic diffusion tractography

**Seed regions.** ROIs were manually segmented on the MNI152\_T1\_1 mm template using the atlas of Talairach and Tournoux (1988) to produce an anatomically correct segmentation. Segmentations were then compared with the FMRIB58\_FA\_1 mm template to ensure an anatomically-robust match. The goal of these segmentations was to be exclusive of any ambiguous voxels, rather than inclusive of all voxels in a particular region of interest. All standard space ROIs were transformed into each subject's native space using the inverse registration matrix described above. We did not know how the plane of acquisition and/or the dominant white matter trajectory in the corpus callosum might influence perceived tract location, so we included three cortical areas chosen for their anatomic properties in relation to the sagittal plane: primary motor cortex (PMC) is largely perpendicular to the midline in the coronal plane; supplementary motor area (SMA) parallels the midline;



**Fig. 1.** A Method for Improving the Colocalization of Tractography.

Fig. 1 – Probabilistic tractography generated in a subject's native space must be registered into a common space for subsequent comparison. Three methods for tract registration (A, B, and C) are illustrated, separated by hashed grey lines; the final tractographic product of each method is highlighted in an oval. These three final products are the basis for comparisons throughout this report: the standard, automated, and manual registration methods. The standard method (A) utilized whole-brain affine and non-linear registration of subject fractional anisotropy images to the FMRIB58\_FA template to produce standard space tractography (oval, A). As illustrated elsewhere in this text, whole-brain registration does not align the highest-value voxels, compromising subsequent quantitative comparisons. A more accurate alignment can be achieved by registering single tracts to a more specific target: the average of all similarly-generated tracts (first the average of tracts registered through the standard method [A, \*] and subsequently to the more tightly-clustered average of colocalized tracts [B,  $\alpha$ ]). Applying an amplitude threshold to the average tract isolated the core tract architecture, further improving the target of registration. Our automated approach (B) aligned standard space tracts to the average tract through two iterations of affine (xyz-translation only, to avoid warp and twist) registration, averaging, and re-registration [steps 1 and 2]. A minority of tracts were further-improved by registration using weighting volumes; we identified the registration method (with or without weighting; two amplitude thresholds; six cost functions) that yielded the smallest tract mismatch through an automated selection algorithm. The final colocalization matrix was merged with the whole-brain registration matrix from A to produce a single warp matrix [step 3], which was then applied to native space tracts [step 4]. This yielded tracts that were registered to a common “tract space” via a single warp step (oval, B). Finally, visual inspection of each automatically-colocalized tract and manual correction of persistent mismatches (C) provided a “gold standard” for comparing tract colocalization. Manually colocalized tracts (oval, C) were compared with tracts registered through the standard registration method (oval, A) in subsequent quantitative analyses. The average of these manually-colocalized tracts (C,  $\Omega$ ) served as the gold standard for location assessments – the best estimate of tract location when mismatches are minimized. The Linux application performing these steps can be reviewed in [Supplemental Data 1](#).

and visual cortex (VC) is largely perpendicular to the midline in the horizontal plane. The PMC segmentation utilized previously-defined anatomical landmarks (Chouinard and Paus, 2006; Picard and Strick, 2001; White et al., 1997). The VC segmentation combined Brodmann Areas 17 and 18, as the projections from VC that reach the corpus callosum originate at the margin of these Areas (Innocenti et al., 1977; Putnam et al., 2010). The corpus callosum segmentation was three voxels wide, centered on the sagittal plane, and included the full anterior-posterior extent of the callosum. To prevent tracts from crossing the midline in non-anatomical fashion (i.e., through CSF or dura), we created an exclusion mask by segmenting the midline sagittal plane, 1 mm wide, with cutouts for all midline white and grey matter structures. This mask terminated any non-anatomical streamlines but permitted passage of interhemispheric, anatomically-plausible streamlines.

It is notable that each of the cortico-callosal projections produces a relatively-simple U-shaped tract. Our new tract-to-tract registration

method (limited to xyz translation, see below) of these highly-similar structures performed very well. In addition to evaluating these cortical tracts, we wished to determine whether (1) location mismatch affected smaller, more anatomically constrained, and phylogenetically older tracts (subcortical) to the same extent that it affected cortico-callosal tracts and (2) our colocalization method was effective for tracts with relatively-complex geometry compared with the U-shape of cortico-callosal tracts. To test these questions we applied our colocalization method to the thalamus-putamen (thalamo-putaminal) and globus pallidus-red nucleus (pallido-rubral) connectivity, which we have studied previously using probabilistic tractography (Blood et al., 2012). The thalamic segmentation included all thalamic nuclei. The pallidum segmentation included pars interna, pars externa, and the intrapallidal laminae. Descriptions of how we aimed to isolate the desired tracts are described below in **Probabilistic diffusion tractography**. All segmentations can be visualized in [Supplemental Fig. 1](#).



**Probabilistic diffusion tractography**, (henceforth named as “tractography”) was generated using the output of *bedpostx* for each subject in native space using *probtrackx*. Left and Right ROIs were run separately. Single cortical ROIs were set as seeds for tractography, with the corpus callosum ROI set as a non-terminating waypoint (which preserves only those tracts that extend from a seed voxel to any voxel within the waypoint ROI); hence, the goal was to extract interhemispheric connectivity for cortical seed regions. Streamlines that reached the sagittal exclusion mask (i.e., non-anatomical streamlines) were rejected. Subcortical tractography utilized a combination of waypoint and exclusion masks to isolate the ipsilateral afferent tracts. The thalamo-putaminal projection was selected using the thalamus as seed, putamen as waypoint, and the pallidum as an exclusion mask to minimize striato-pallido-thalamic projections. Thalamo-putaminal tractography also utilized a full-plane midline sagittal exclusion mask to assure that only ipsilateral tracts were included. The pallido-rubral projection utilized the pallidum as seed, ipsilateral red nucleus as a waypoint, and the contralateral superior cerebellar peduncle as an exclusion mask to eliminate the ascending cerebellothalamic tract coursing through and directly adjacent to the red nucleus (Gallay et al., 2008). Default parameters for running *probtrackx* were utilized: mode = network; curvature threshold = 0.2; steps = 2000; samples = 5000; loopcheck engaged.

**Thresholding tractography to increase the precision of individual tract location and to facilitate alignment of tracts across individuals.** In this study, thresholding was an essential step in defining the core, most-relevant voxels, for individual tracts. Thresholding was also essential for improving alignment of tracts by ensuring that registration compared the most-relevant voxels of individual and target tracts. However, different amplitude thresholds were necessary to optimize these two parallel goals. These differing threshold requirements reflect the different sources of dispersion in streamline location: for individual tracts, low-frequency eccentric streamline placement is an intrinsic feature of probabilistic tractography computation; for group averages (multiple 3-dimensional tract volumes averaged at each voxel), individual or experimental factors that led the entire streamline bundle to be eccentrically placed lead to location dispersion. Below we describe the purpose and the procedures of thresholding in each of these two cases.

Probabilistic tractography produces streamlines that reach virtually every voxel in the brain, albeit with most voxels accounting for a negligible number of streamlines. These lowest-amplitude voxels are non-specific; any given voxel is likely to be contacted by streamlines from most seed regions, and in most individuals. However, because the streamlines at these non-specific voxels were sparse compared to voxels near the center of the tract, we were able to use thresholding at the individual level to exclude these non-specific voxels from our analyses. Specifically, as a means of isolating the core of individual tracts, we applied amplitude thresholding (akin to a high-pass filter) to reduce this “noise” and focus our signal on the core of each tract (Ciccarelli et al., 2006; Guye et al., 2003; Hu et al., 2011). For all individual tracts we applied a threshold to retain voxels in the upper-most 90% of maximum amplitude; these threshold-limited tracts were utilized for all subsequent steps that involved quantification of tract amplitude, including group comparisons.

For group averages of tractography, the dispersion of streamlines resulting from individual differences in the location of the core tract also results in dispersion of streamline intensity across a greater number of voxels; individuals whose tract location deviates substantially from the group produce a dilution of amplitude at the center of mass. As a means of providing iteratively more localized (as opposed to dispersed) template “average tracts,” we applied amplitude thresholding (akin to a high-pass filter) to increasingly hone in on the region with the highest probability of tract location across individuals. This iterative improvement in the target of registration – the core of the average tract – allowed each tract to be aligned to produce maximal inter-subject or inter-session tract overlap. Averaged tracts required a higher amplitude threshold to isolate the core tract, as the large majority of individual tracts have a sharp cross-

sectional contour (values drop rapidly from the peak) but average tracts invariably have a more rounded cross-sectional contour and more regular peak height along the length of the tract. The goal of these thresholds was to isolate the essential tract architecture (the common U-shape) at the center of mass of the average tract to serve as a target for registration; these higher thresholds were not used for quantification or between-subjects comparisons, only for refining the group averages used as a target for registration steps. For each subsequent utilization of average tracts, we will specify in the corresponding Methods section the optimal threshold to isolate the core tract. While lower-amplitude voxels may be physiologically relevant – changes to the less-robust connections between regions may still have critical structural and functional implications – these less-robust connections require anatomical selection methods that were not the focus of this paper.

## 2.5. Registering tracts into standard space

Colocalizing tracts – that is, bringing tractography from all subjects into a shared space before groupwise comparisons to minimize location mismatch – is essential for comparing tract metrics such as tract amplitude, volume, and length.

### 2.5.1. Whole-brain registration: the standard registration method (Fig. 1A)

Native space tractography was spatially normalized to the FMRIB58\_FA template with *applywarp*, using the native-to-standard matrix derived from whole-brain native to standard space non-linear registration (see section 2.3). This standard method for registering tracts from native to standard space assumes that tract location mirrors surrounding grey and white matter structures. The goal of this project was to improve on this whole-brain registration approach.

### 2.5.2. Tract-to-tract registration: automated colocalization (Fig. 1B)

*Improving the alignment of individual tracts requires a better approach to registration: direct tract-to-tract registration.* We wished to improve tract registration by aligning individual tracts, rather than by aligning grey and white matter structures; we did this by registering each tract to the average tract. To create this improved target for registration, we averaged the standard space tracts for a particular cortico-callosal pairing, as detailed in [Inline Supplemental Material 1](#). This average tract (Fig. 1A, \*) then served as the target for tract-to-tract registration, now using only linear xyz translation, without rotation, stretching, skew, or non-linear translation. Though we trialed non-linear modes of registration, we found that these frequently distorted the native tract architecture and led to non-anatomic displacements; the more-limited xyz translation simply moves the tract in space without warping, producing a more reliable tract-to-average registration.

In order for the average tract to serve as a superior target for registration, it should be limited to those voxels most likely to be shared across the group of tracts; accurate registration requires a precise definition of the desired target. We empirically noted that using normalized and thresholded versions of tracts and average tracts improved the accuracy of registration, presumably by defining the core tract structure and reducing inter-subject variability in maximum tract intensity ([Inline Supplemental Material 2](#)). Note, however, that we used normalized versions of the tracts only to define each individual's registration matrices; the tractograms that were registered, and subsequent comparisons of tract amplitude and location in these images, were always performed on non-normalized tracts thresholded to retain the uppermost 90% of tract amplitude. By iteratively (1) registering tract-to-average tract, (2) creating a new average from the newly registered (more accurate) tracts (Fig. 1B,  $\alpha$ ), and then (3) registering again to the new, more accurate average, we progressively improved tract alignment. Two rounds of linear registration to the average tract optimized the colocalization of tracts; trials of three or more rounds of registration produced only marginal improvement over two rounds.

For approximately one quarter of the 222 cortico-callosal tracts

registered through the automated method, simple affine registration left tracts with >1 mm separation in center of gravity from the target. For those outlier tracts, we carried out an additional automated affine registration step that utilized weighting volumes (Jenkinson, 2014) to improve tract-to-average match. We used the normalized average, thresholded to retain the uppermost 90% of voxels, as the reference weighting volume, and the normalized individual tract as the input weighting volume. All outlier tracts were registered using the following combination of parameters: each of six FSL-provided cost functions (mutual information, normalized mutual information, normalized correlation, label difference, correlation ratio, and least squares); two tract thresholds (retaining either the uppermost 90% or 75% of voxels). We identified (via a uniformly-applied algorithm that select the smallest root-mean-square distance) the registration that minimized tract mismatch from among these twelve parameter combinations. Note that every combination of cost function and tract threshold was optimal for at least one tract, with no clear predictive pattern of which combination would be effective; trying all combinations and algorithmically identifying the closest match optimized the Automated registration. Likewise, for approximately 10% of subjects, weighed registration led to markedly inaccurate placement (off by tens of millimeters), far less accurate than our standard Automated method (Fig. 1B1); in these unusual circumstances, the standard Automated method (no weighting) was preserved. We suggest that users register all subjects using weighting volumes, even those with minimal residual separation, as even small improvements boost accuracy. However, tracts registered within 1.25 mm of the target COG were unlikely to be improved by weighting volumes registration; users may speed registration by increasing the mismatch threshold, set by default at 1.25 mm. As with any registration method, visual inspection following Automated colocalization was an essential quality control measure.

Finally, the resulting warp matrices from whole-brain registration (native to standard space) and colocalization (standard space tract to average tract) were combined to produce a final registration matrix (a summation of each registration step). We then applied this comprehensive registration matrix to native space tractography in a single step (Fig. 1B4; [Inline Supplemental Material 3](#)). Note that this method combined all registration steps into a single warp from native to tract space; all tracts subsequently utilized for quantification thus had undergone a single, comprehensive registration step and were not subjected to iterative registration transformations, which can lead to cumulative distortions. We then applied this comprehensive registration matrix to native-space tracts to colocalize the tracts in a single registration step: `applywarp -ref = FMRIB58_FA.1 mm.nii.gz -i native_space_tract.nii.gz -warp = comprehensive_colocalization_matrix -rel -out = colocalized_standard_space_tract.nii.gz`.

### 2.5.3. Tract-to-tract registration: manual colocalization (Fig. 1C)

Correcting tract location by direct visualization and manual adjustment allowed us to define a “gold standard” for registration. Following automated colocalization, each tract was visually compared with the colocalized average using `fslview` ([Inline Supplemental Material 4](#)). Tracts were viewed in sagittal, coronal, and axial planes to identify mismatches between single tracts and the average tract. Detection of mismatches was facilitated by viewing tracts at a range of high thresholds, often restricting tracts to only a few voxels per plane. The peak value of individual tracts and the contour of the tract were equally important in determining the accuracy of alignment, making it essential to assess tracts across a range of thresholds, not limiting one’s assessment to the tract peak. For tracts whose contour did not readily match the average tract (e.g., tracts with bifid peaks joined by a saddle region, or tracts whose cross-sectional area was similar to a comma, rather than a bullseye), tracts were positioned to array the amplitude symmetrically around the average tract. Manual registration was achieved by editing the registration matrix to adjust the xyz position (no twist, warp, or stretch) and repeating `convertwarp` and `applywarp`, as described above. We generated an average of these manually-colocalized tracts (for each

cortico-callosal pairing and from all 37 subjects; detailed in [Inline Supplemental Material 1](#)); this average of the manually-colocalized tracts served for subsequent steps as the standard for assessing colocalization (Fig. 1C,  $\Omega$ ). We extracted the peak value of this average tract (37 subjects, for manually colocalized and whole-brain registered conditions) at the corpus callosum to quantify the amplitude lost to inaccurate registration. The average of these manually colocalized tracts served as a “gold standard” for tract location; individual tract location was compared with this manually-colocalized average to assess degree of mismatch.

## 2.6. Assessment of tract amplitude distribution and the degree of location mismatch

We wished to identify the distribution of tract amplitude around the tract peak, i.e., how quickly streamline counts drop as one moves away from the tract peak. Quantifying this rate of decline in single tracts allowed us to gauge the importance of tract location in amplitude assessments – wide distributions are more permissive of mismatch, while tract mismatch may have great impact on narrow distributions. Next, we quantified the degree of location mismatch as a function of amplitude; we quantified this colocalization at discrete thresholds from low-to high-amplitude. Specifically, as described in sections 2.4 and 2.5.2, most voxels contacted by probabilistic streamlines are non-specific and have very low amplitude; thresholding allowed us to distinguish the impacts of mismatch on the tract core (the portion of each tract containing the highest density of streamlines) from the impacts of mismatch on non-specific voxels (Ciccarelli et al., 2006; Guye et al., 2003; Hu et al., 2011). Finally, we assessed the degree of location mismatch between tracts through two methods: the volume of overlap between tracts, and the separation in center of gravity between tracts.

### 2.6.1. Single-tract amplitude distribution

Using `fslstats`, we determined the amplitude (streamline count) and coordinates of the peak midline voxel for each tract. Each tract was visually inspected to assure that the selected peak voxel was not a “stray” high-amplitude voxel; all fell at the tract center. We then generated masks using `fslmaths` of voxels at  $\pm 1$ , 2, and 3 mm from the peak to extract the amplitude from each of these surrounding voxels using `fslstats`. Thus, we extracted amplitude from spherical surfaces centered on the peak voxel; these spheres had radii of 1, 2, and 3 mm.

### 2.6.2. Distribution of amplitude within a group

Quantifying tract mismatch allowed us to compare registration techniques and define goals for colocalization. We assessed the degree of anatomical colocalization in cortico-callosal tracts for all registration methods (whole-brain registration, the standard method; automated colocalization; and manual colocalization) using two discrete approaches. First, in order to assess the overlap of single tracts to the gold standard, we calculated the Dice similarity coefficient (DSC) for each tract relative to the average of manually-colocalized tracts. All tracts and averages were thresholded at discrete percentages of the maximum value at the corpus callosum (to retain voxels with the uppermost 25%, 50%, 90%, and 99% of maximum amplitude) and then binarized to produce a mask image. These mask images were used only to quantify tract overlap, and were not used in later measures of tract amplitude. Volumes of each mask and the overlap between a tract mask and the average mask were used to calculate the DSC as follows:

$$2 \times \frac{(\text{Volume}_{\text{Individual tract}} \cap \text{Volume}_{\text{Average tract}})}{\text{Volume}_{\text{Individual tract}} + \text{Volume}_{\text{Average tract}}}$$

Second, we compared the center of gravity, a weighted average of signal intensity and location, between individual tracts and that of the manually-colocalized average tract ([Inline Supplemental Material 5](#)).

Since cortico-callosal tracts vary in the location of their maximum value (e.g., cortex, corona radiata, or corpus callosum), center of gravity

measures for the whole tract differ substantially between individuals, independent of registration method. By restricting assessment of center of gravity to the value at the midline sagittal plane (using the corpus callosum segmentation), where the direction of travel was orthogonal to the plane of measurement, we reduced the impact of individual variance in tract morphology and instead focused the center of gravity assessment on the impact of tract location. Center of gravity assessment was achieved using *fslstats*. Finally, the root-mean-square distance between an individual tract's center of gravity and that of the average tract was calculated for each registration method (standard, automated, and manual). For subcortical tracts, center of gravity measures were calculated in similar fashion, but utilized a distinct mask volume to extract individual center of gravity measures: a volume centered on the maximum value of the average tract. Each of these subcortical mask volumes was  $3 \times 7 \times 9$  mm (x-y-z), dimensions selected to parallel those of the sagittal corpus callosum mask used to measure cortical tracts: 3 mm, the width of the callosal segmentation; 7 mm, the maximum superior-inferior dimension of body of the corpus callosum mask; and 9 mm, 3 times the mean variability in tract location for cortico-callosal tracts. The masks were oriented predominantly in the Y plane and were centered at (MNI coordinates, mm): left thalamo-putaminal tract, -26 -28 -2; right thalamo-putaminal tract, 29–25 -3; left pallido-rubral tract, -5 -10 -11; right pallido-rubral tract, 6–10 -10.

## 2.7. The effect of location mismatch on group comparisons of tract amplitude

We used two modes of quantification to study the effects of tract location on measures of tract amplitude: (1) in single subjects scanned at two time points, and (2) between subjects in artificially-constructed groups. Tracts were quantitatively assessed in either a simple group average method (mean streamline counts for within-subject repeated scans) or through voxelwise non-parametric permutation testing using the FSL tool *randomise* (for between-groups, single-scan comparisons). Note that permutation tests do not easily accommodate repeated-measures datasets, as these violate the principle of null-hypothesis exchangeability, though recent updates to *randomise* hold promise for repeated-measures testing in tractography (Behrens T, 2014); for this reason, *randomise* was not utilized to assess repeated scans.

### 2.7.1. Comparison of averaged tracts: assessment of tractography from repeated scans

Comparing scans from two time points in individual healthy subjects allowed us to quantify the impact of tract mismatch on estimates of tract amplitude. When using tractography to study white matter changes over time, such as in response to disease treatment, or tracking disease progression over time, an implicit assumption is that control or placebo-treated subjects will have minimal changes in tract amplitude (streamline count or density) over short time intervals. However, if an individual's core tract occupies different locations in repeated scans, comparison of tract amplitude across scans may be inaccurate. An average of two such tracts from one individual integrates differences in tract amplitude and tract location between scan sessions. If the two tracts were perfectly colocalized, the amplitude of this average tract would equal the mean amplitude of the two tracts, independent of location:  $\text{Amplitude of Average Tract}_{1,2} = \text{Mean}(\text{Amplitude Tract}_1 + \text{Amplitude Tract}_2)$ . A reduction in the amplitude of the average tract relative to the location-free ideal (the mean amplitude of the two tracts considered separately) suggests that location mismatches between scan 1 and scan 2 led to an under-estimate of the true amplitude of tracts 1 and 2. If such inaccuracies occur in healthy individuals scanned at short time intervals, it appears likely that tracts averaged across a group of individuals will suffer from similar or greater reductions in estimated tract amplitude.

We set out to quantify the impact of within-subject mismatches on tract amplitude in healthy volunteers scanned at two time points, one month apart. Over this time interval we expected no gross changes in the

amplitude of tractography. For the 14 subjects with repeated scans, we first quantified each individual's tract amplitude (mean streamline count), independent of location, for Tract\_1 and Tract\_2 (tracts derived from the two scans - [Inline Supplemental Material 6](#)). We then averaged those two values to give a location-free estimate of tract amplitude across the two scans. Second, we averaged Tract\_1 and Tract\_2 for each individual and then quantified the amplitude of this averaged tract for each individual. These two approaches (extract value from each tractogram, then average value; average tractograms together, then extract value) allowed us to separate inter-scan differences in amplitude from inter-scan differences in location. We compared the location-free amplitude estimate with the location-incorporating amplitude estimate for each individual for three conditions: for tracts registered via the traditional, whole-brain method, on tracts colocalized through our automated registration method, and for tracts that were manually colocalized. The ratio of location-incorporating:location-free amplitude for each individual was assessed across the group of 14 subjects to give a group estimate of the impact of location mismatches on amplitude measures. Ratios were compared between tracts aligned by the standard, whole-brain registration and by tracts aligned by manual colocalization using a two-tailed, paired-samples *t*-test.

### 2.7.2. Voxelwise comparisons: defining artificial groups

We defined artificial groups (demographically matched into pairs of individuals, then sorted into non-overlapping groups) to simulate the group-wise matching employed to study disease or treatment (e.g., patients vs. controls, active drug vs. placebo). We recognized that while mismatches in tract location might compromise the accuracy of amplitude measurement, these effects would likely have disparate impacts under two circumstances: first, if Group A and Group B have similar mean tract amplitude, location mismatch could produce false positive results; second, if Group A and Group B have very different mean tract amplitude, location mismatch could produce false negative results. It is important to be explicit about how we define “mean tract amplitude.” For each individual subject assessed in her own “tract space,” the average number of streamlines per voxel across a tract is a pure metric of amplitude, independent of tract location; this extracted measure estimates what an individual tract would contribute to a group comparison if inter-subject differences in tract location were ignored. If one averages these individual measures of tract amplitude across a group, the resulting mean tract amplitude is an estimate of that group's location-independent tract amplitude. Note that the description of location-free amplitude in section 2.7.1 is identical to the mean tract amplitude. In contrast, averaging 3-dimensional tracts in standard imaging space across a group (now speaking of voxelwise averaging of tractograms, not extracted data) integrates amplitude and location information. We used the location-independent measure, mean tract amplitude, to define our artificial groups. By designing groups that were matched for mean tract amplitude, and separate groups that were discordant for mean tract amplitude, we were therefore able to identify the impact of tract location on false-positive and false-negative voxelwise results, respectively.

Specifically, from the 37 healthy subjects in this series, we identified pairs of subjects that could be matched 1-to-1 for gender, handedness, scanner and age (within 6 years, mean difference: 2.75 years). This produced 16 pairs of healthy subjects that were used for subsequent quantitative assessment (see [Supplemental Fig. 2](#) for a schematized version of this process). The remaining 5 subjects had no match. These groups allowed us to test the hypothesis that quantitative assessments of tractography are influenced by both the amplitude and location of individual tracts, and, by corollary, that reducing variance in tract location improves the ability to identify true differences in tract amplitude. These 16 matched pairs were used to define artificial groups, with each group containing 32 healthy individuals (16 vs. 16, described in the following paragraph and in [Supplemental Fig. 2](#)). Each of these 16 matched pairs was divided between distinct subgroups of 16 individuals and differed from other artificial groups only in how matched pairs were sorted



between the two subgroups.

In our group of 37 healthy subjects, the left primary motor cortex seed consistently produced robust tractography, so we used this tract to define the mean tract amplitude for each individual. Therefore, only left primary motor cortex tractography was assessed for 16 vs. 16 artificial groups. The mean tract amplitude for the 32 members of this cohort followed a Gaussian distribution (Supplemental Fig. 2), but the two members within a matched pair were often discordant for mean tract amplitude. Flipping the members of a pair between subgroups could therefore be utilized to bias these artificial groups to sample from particular portions of the Gaussian distribution. We flipped subjects within these 16 matched pairs to produce artificial groups that retained demographic matching, but whose tract amplitude was either: balanced for mean tract amplitude (<2% difference in mean); overtly mismatched (>25% difference in mean); or intermediate mismatched (9–15% difference in mean - Supplemental Fig. 2). Since our explicit goal was to identify the impact of spatial dispersion on subsequent voxelwise testing, when assembling these groups we arranged pairs to maximize differences in tract location. Note that maximizing location mismatch came last in priority when arranging matches, after assuring that groups were of similar mean age and after segregating by amplitude. For each type of comparison (balanced, mismatched, and intermediate), three sets of cohorts were arranged, for a total of nine matched cohorts. Given that these were artificially-generated groups, we produced three cohorts for each comparison to ensure that the findings were not unique to one artificial pairing of subjects. Each cohort of 16 vs. 16 maintained 1-to-1 demographic matching, and the largest difference in mean age between groups was <2 years. The degree of location mismatch between groups in a cohort averaged 1.1 mm (range: 0.05–1.7 mm).

### 2.7.3. Voxelwise comparisons: *randomise* comparisons for artificial groups

For each cohort of 16 vs. 16 subjects described in section 2.7.2 (nine total cohorts, three each of Balanced, Mismatched, and Intermediate), each subject's tracts were registered in one of two ways: by the standard, whole-brain registration method; or by manual colocalization. Between these two conditions, the subjects, tracts, and individual amplitudes were identical: they differed only slightly in tract location (mean difference: 1.97 mm), having been registered by the standard whole-brain method or colocalized to the average tract. The fundamental premise of our method is that such slight differences in location can have large impacts on measures of tract amplitude; depending on the relative location-free tract amplitude of the two groups being compared, such mismatches could yield either false-positive or false-negative results. Each cohort was assessed using the FSL tool *randomise*, with standard registration and manual colocalization conditions run in parallel (see section 2.10 for a detailed description of the *randomise* options utilized).

As these *randomise* comparisons differed only in the degree of tract colocalization, but were otherwise identical in individual tract amplitude and relative architecture, voxelwise output from *randomise* was compared between runs to determine the effect of colocalization on voxelwise testing. For each *randomise* comparison, we first calculated the cluster mass, a metric reported to be more sensitive than comparisons based exclusively on extent or amplitude of clusters (Bullmore et al., 1999; Hayasaka and Nichols, 2004).

For all suprathreshold voxels ( $t > 2.33$ ) in the uncorrected *randomise* maps, we extracted cluster volume and mean amplitude, and multiplied volume X amplitude to calculate the cluster mass. Given that cluster mass was extracted from discrete statistical tests and was not a direct statistical comparison of those discrete tests, and the goal of this analysis was simply to demonstrate that reducing location mismatch improved the accuracy of voxelwise testing, these are presented as simple ratios of cluster mass output (whole-brain registration: manual colocalization). Thus, the relative cluster mass values (which differed markedly between whole-brain and colocalization methods of registration) may be seen as a qualitative, not quantitative, means of assessing the impact of colocalization on the accuracy of voxelwise testing. The results of *randomise*

testing were similar within each type of comparison (the three balanced cohorts all produced similar results, the three overtly mismatched cohorts all produced similar results, etc.), so results for two groups from each comparison type are presented here.

We recognize that some readers may wish to gauge the impact of colocalization on the statistical significance of voxelwise testing, when corrected for N comparisons across the volume of interest. Therefore, we utilized cluster-based thresholding to perform family-wise error (FWE) correction for multiple comparisons, using the null distribution of the maximum cluster mass (the *randomise* “-C” flag, threshold  $t > 2.33$ , Behrens et al., 2014). Please note that the cluster mass we calculated using the uncorrected t-statistic maps (paragraph immediately above) is distinct from this FWE correction within *randomise*; while they share terminology and t-threshold, they are otherwise completely distinct means of assessing group differences.

## 2.8. Identifying sources of location mismatch

*Variability in tract location may reflect the normal spectrum of individual differences, but may also reflect remediable experiment-related factors (e.g., movement, partial volume effects, B-field inhomogeneities).* To test whether there was a relationship between head motion and tract location, we used linear regressions to assess both DSC and tract displacement (center of gravity, single tract vs. average tract) measures relative to the absolute head displacement during scan acquisition, a motion correction metric from the prior *eddy\_correct* step (section 2.3, Preprocessing). If one assumes that tract location is stable within an individual over short periods of time, subjects scanned twice (one month separation between scans) offer the opportunity to distinguish between individual and experimental contributions to location mismatch. Note that the duration of intra-subject imaging stability is not defined, with some authors utilizing separations as short as 1 h (Landman et al., 2011); we chose to rescan following a one month delay to more-closely model the time intervals commonly used to assess clinical treatment. For each subject that was scanned twice, we calculated the root mean square distance between individual tracts (scan 1) and either (A) the same individual's tract from scan 2, or (B) the scan 1 tract of a different individual (each of the other 13 subjects, serially compared). The mean of all individual-to-other-individual distances revealed the totality of all causes of location mismatch, while the repeated-scan distance revealed the location mismatch due to experimental factors. The difference between distance A and distance B approximates the location mismatch due to normal differences between individuals.

## 2.9. The effect of location mismatch on mean FA

*Adjusting tract location to reduce mismatch may improve quantification of amplitude, but will also lead to sampling of a distinct diffusion microenvironment.* Researchers may wish to present both traditional diffusion metrics and quantified tractography. Whether to use traditionally-registered or colocalized tracts to extract data is uncertain. We asked whether altering the precise location of tracts (colocalization) led to predictable distortions in measures of FA. We used tracts in standard space, thresholded to retain the uppermost 25% of voxels, to extract mean FA values for each subject. We measured mean FA under two conditions, using either the whole-brain registered or manually colocalized tract.

## 2.10. Statistical assessment

We assessed interscan location mismatch (center of gravity and DSC, section 2.6.2), mean tract amplitude (section 2.7.2), and age in the pool of 37 subjects using the Wilcoxon-Mann-Whitney test, as these data were not normally distributed. We assessed these same variables for the 14 subjects with two scans using paired-samples, two-tailed t-tests. Likewise, we assessed mean FA comparisons (section 2.9) using paired-



samples, two-tailed t-tests. The requirements for utilizing paired-samples testing were confirmed (Armitage and Berry, 1987; Ha and Ha, 2011; Swinscow and Campbell, 1997). Results for each of these statistical comparisons were corrected for multiple comparisons by dividing the significance threshold by six (left and right hemispheres, three seed regions;  $p_{\text{corrected}} = 0.05/6 = 8.3 \times 10^{-3}$ ). We investigated the relationship of head motion with DSC or tract displacement using linear regression. For the left primary motor tract, the only tract utilized for artificial group comparisons, we calculated the Pearson's correlation coefficient between the amplitude of tract mismatch (the root mean square distance between a tract's center of gravity and the center of gravity of the average tract) and the mean value of that individual's core tract (the mean of all voxels with value > 75% of the max tract value). Statistical comparisons were conducted using Stata (StataCorp LP, 2013 Release, College Station, TX).

Permutation-based nonparametric, voxelwise assessment of groups of tracts was conducted using the FSL tool *randomise*. Details of how groups were constructed for *randomise* comparisons are included in section 2.7.2. For all comparisons, *randomise* was run with 5000 permutations, variance smoothing of 8 mm, cluster mass t-threshold of 2.33, and masked using the group average tract (all 37 subjects, thresholded to include the upper 90% of voxels).

### 2.11. Data availability

The code utilized for tract-to-tract registration is supplied as a supplemental data file. The authors will share raw tractograms and/or extracted tract-based data with interested parties following a request to the corresponding author.

## 3. Results

We assessed tractography derived from six independent cortical seed regions (left and right hemispheres, for primary motor cortex, supplementary motor area, and visual cortex) constrained by a midline mask of the corpus callosum. All tracts followed a common U-shaped trajectory

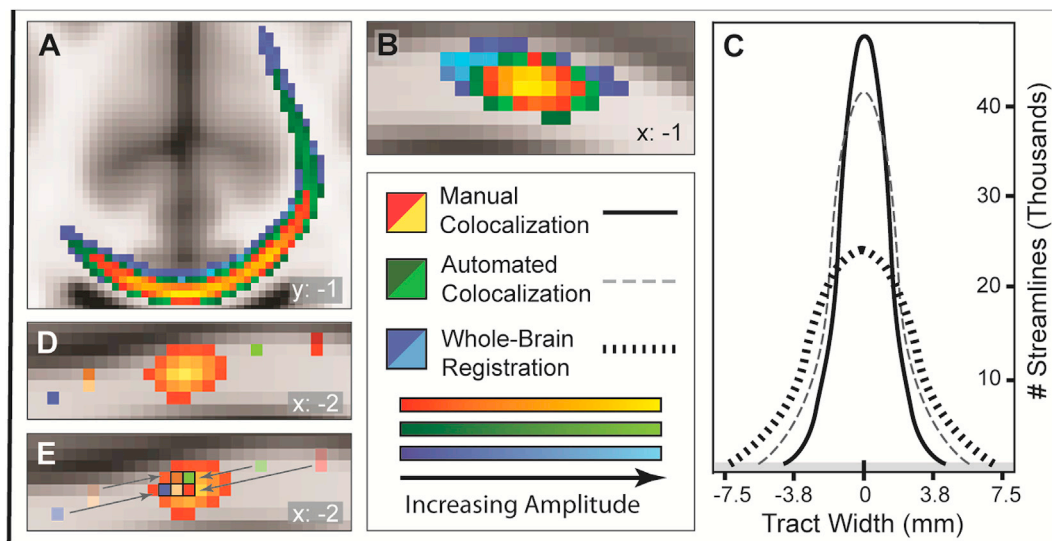
through the corpus callosum (Fig. 2A), and pierced the midline sagittal plane at near-perpendicularity. This made inter-subject comparison of trans-callosal tract location more straightforward than for tracts with complex architectures or divergent projections. To determine whether location mismatch was a general feature of all probabilistic tractography, we assessed tracts derived from four independent subcortical seed-target pairs to identify the left and right putamino-thalamic, and the left and right pallido-rubral tracts. We quantified inter- and intra-subject variables that contributed to poor tract colocalization. Finally, we assessed the degree to which tract amplitude was underestimated when tracts were assembled into groups, a consequence of poor tract colocalization (aka, location mismatch).

### 3.1. Dispersion of amplitude in single tracts

Tract amplitude, a count of the number of streamlines transiting a particular voxel, was arrayed around the tract peak in a regular pattern that was similar across seed regions and individuals (Table 1). For the first millimeter moved from the tract peak, amplitude fell by 21% from the peak value. From 1 to 2 mm, residual tract amplitude fell further by 38%, and from 2 to 3 mm, residual tract amplitude fell further by 46%. At 3 mm from the tract center, therefore, tract amplitude had fallen by a cumulative 74% from the tract peak ( $21\% + [38\% \text{ of } 79\%] + [46\% \text{ of } 49\%] = 74\%$ ). The orientation of the seed region relative to the corpus callosum, and the dominant plane in which the tract traveled to reach the corpus callosum, had no impact on the rate of dispersion in tract amplitude.

### 3.2. Assessment of colocalization following the standard whole-brain registration method

The registration of native space whole-brain volumes into standard space, and the subsequent use of that registration matrix to warp native space diffusion tractography into standard space (Fig. 1A), produced poor colocalization of cortico-callosal tracts (Fig. 2). While overlap of a



**Fig. 2.** Whole-brain registration produces tracts that overlap poorly.

Fig. 2 – The application of whole-brain registration matrices to tractography produced standard space tracts that did not share a common space. While whole-brain registration and averaging of tracts produced a seemingly-tight bundle (blue, representative sample from the right supplementary motor area in coronal, A, and sagittal sections, B, thresholded at 50% of maximum amplitude, from 37 healthy controls), overlay of colocalized tracts (green, red-yellow) demonstrated substantial opportunities for improved precision. In C, a representational plot of data from panel B, increasing accuracy of registration was associated with higher peak values and a narrowed distribution. In D, the 10% of tracts with the greatest spatial dispersion (aligned using the standard registration method, peak values represented by blue, tan, green, or deep-red voxels) are superimposed on the manually colocalized average tract (right primary motor cortex, thresholded at 50% of maximum amplitude) to illustrate the degree of location mismatch that is common with whole-brain registration. Alignment of all tracts with their average tract markedly reduced location mismatch and improved the opportunity for voxel-wise comparisons, as illustrated for those 10% of tracts with the largest location mismatch (E, illustrating the change from panel D).

**Table 1**  
Amplitude dispersion by millimeters moved from tract center, 37 controls.

A. Each mm moved away from the tract center leads to an ADDITIONAL loss of value by this amount:							
	Left-PMC	Right-PMC	Left-SMA	Right-SMA	Left-VC	Right-VC	Mean for all ROIs
Moving from 0 to 1mm	21.9%	21.3%	19.2%	20.3%	23.0%	22.7%	21.4%
Moving from 1 to 2mm	37.7%	37.6%	32.3%	38.3%	40.1%	40.1%	37.7%
Moving from 2 to 3mm	46.5%	47.9%	38.7%	46.8%	47.9%	48.5%	46.1%
B. CUMULATIVE loss of tract amplitude at 3mm							
At 3 mm from tract center, tracts are reduced from 100% to this value:	26.0% of peak value	25.6% of peak value	33.5% of peak value	26.2% of peak value	24.0% of peak value	23.9% of peak value	26.7% of peak value

Table 1 Legend: For every millimeter moved orthogonal to the tract axis, tract amplitude decreases by the listed percentage. This percentage indicates the additional loss of value with each mm moved from the tract center, rather than the cumulative total. The mean separation between a given tract and the average was >3 mm for all cortico-callosal tracts, suggesting that poor tract alignment can substantially confound amplitude comparison between tracts. Abbreviations: PMC – Primary Motor Cortex; SMA – supplemental motor area; VC – visual cortex; Thal-Put – thalamo-putaminal; Pall-RN – pallido-rubral.

single tract with the average tract (by Dice similarity coefficient, DSC) was 55–65% at lower-value thresholds (which retained 99% and 90% of voxels, Fig. 3A), at our highest-value threshold (which retained the uppermost 25%, the core of the tract) the overlap was substantially lower, with DSCs of 5.8–18% (Fig. 3B, Table 2). Notably, this degree of tract overlap is highly similar to that accomplished by other tract-to-tract registration methods (Garyfallidis et al., 2015; Olivetti et al., 2016), therein described by the Jaccard index. Voxels within the tract core have a disproportionate impact on comparisons of tract amplitude (Ciccarelli et al., 2006; Guye et al., 2003; Hu et al., 2011). The 6-fold reduction in DSCs for the tract core, relative to the full-threshold tract (Fig. 3A), underscores the fact that improving the overlap of the full-threshold tract is laudable, but is less important than improving overlap of the tract core.

In a parallel measure, the center of gravity in the midline sagittal plane (perpendicular to the primary tract vector) for individual tracts varied substantially from that of the average (mean distances: 2.5–3.9 mm, Fig. 3C, Table 3). This separation might appear small. However, recognition of the rapid fall in amplitude with even small movement away from the tract peak (Table 1) suggests that even these small mismatches produce a comparison of one tract's peak to another tract's slope – while a tract at the center is at maximum value, a tract centered at the mean separation distance (3.2 mm) will represent only 27% of its maximum value. More alarming, the average tract was comprised of individual tracts whose location varied substantially more than the mean separation distance, in some cases by more than 12 mm (Fig. 2D). Furthermore, since tracts were arrayed circumferentially around the center of the mean, tracts on opposite sides of the center had even greater separation from each other: for the two most-separated individuals for each cortico-callosal pairing, the mean separation was 14.3 mm (range: 7.7–20.1 mm). Since tract amplitude fell by an average of 99.6% at 14 mm from tract peak, comparison of tracts at this separation is meaningless.

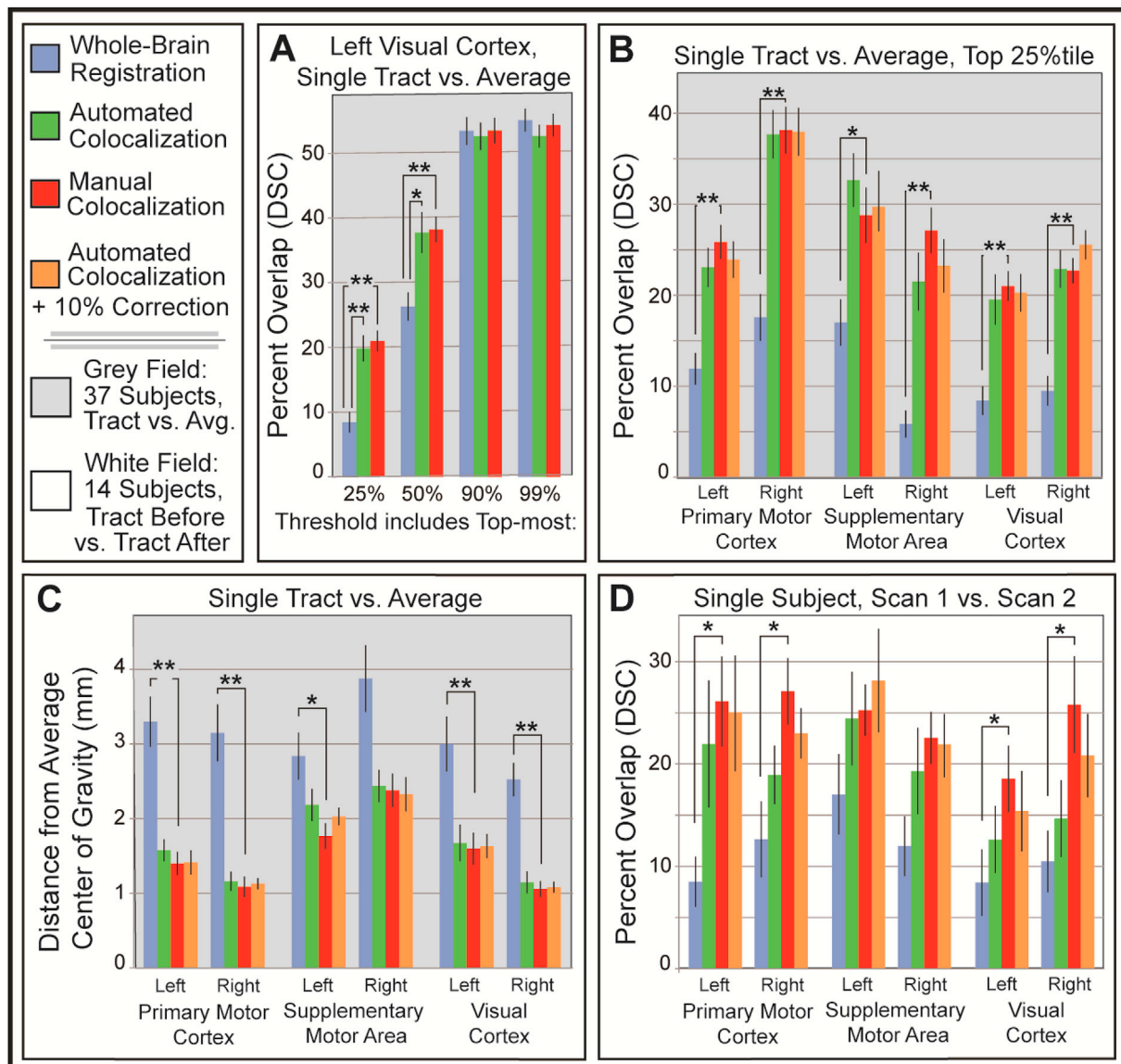
Variability in tract position was most likely to occur along the anterior-posterior axis of the corpus callosum (Fig. 2D) for primary motor cortex and SMA tracts (3- to 4-fold greater than for superior-inferior dispersion), while for visual cortex tracts were dispersed similarly in both anterior-posterior and superior-inferior axes. The magnitude of tract mismatch had no relationship with tract amplitude (Pearson's correlation coefficient = 0.04, 37 subjects).

### 3.3. Assessment of colocalization following automated and manual tract registration

Registration of a tract into a common “tract space” significantly improved the degree of overlap between subjects (Fig. 3A and B, Table 2, DSCs of manual vs. whole-brain registration,  $p$ -values  $< 4.0 \times 10^{-3}$  -  $6.0 \times 10^{-7}$ ), significantly reduced the degree of location mismatch between individuals (Fig. 3C, Table 3, centers of gravity of manual vs. whole-brain registration,  $p$ -values  $< 0.017$ – $6.0 \times 10^{-7}$ ), eliminated extreme outliers in tract position (Fig. 2D and E), and substantially increased the amplitude of the average tract. While manual colocalization (Fig. 1C) produced the greatest improvements in these measures (2.3-fold mean increase in DSC for subject-to-average comparisons; reducing center of gravity difference by 2.1-fold), our automated technique (Fig. 1B) led to substantial improvements as well (1.9–3.7-fold improvement in DSC, and 1.9-fold reduction in center of gravity difference, Fig. 3B–D, Tables 2–4). By supplementing automated colocalization with manual correction of the “worst” subjects (the 10% of tracts with the greatest location mismatch from the average tract), we achieved improvements that paralleled manual segmentation of the full group (Fig. 3B–D). Colocalization boosted the peak value of the average tract (Fig. 2C) for each of the six cortical regions tested, by a mean of 62% (range: 36–81%).

### 3.4. Colocalization of tracts from repeated scans

Whole-brain registration of cortico-callosal tracts in healthy individuals scanned twice (one month separation between scans) yielded tracts with poor overlap (Fig. 3D, Table 2). When comparing scan 1 and scan 2 for individuals registered through the standard method, mean DSCs for the tract core (the uppermost 25% of tract values) were 8.4–17%. Manual colocalization improved the inter-scan overlap of the tract core by 1.5–3.1 fold (mean: 2.1-fold increase in overlap, to DSCs of 19–27%) over standard registration (DSCs of manual vs. whole-brain registration;  $p$ -values  $< 0.024$ – $4.2 \times 10^{-4}$ ). Similar to the effect of colocalization on individual-to-group comparisons, manual correction of the 10% of scans with the worst before-to-after mismatch achieved substantially improved DSC over automated colocalization alone (Fig. 3D). Differences in center of gravity between scans 1 and 2 reduced by 7.7–52% following colocalization (Table 3), but these improvements did not pass our significance threshold, corrected for multiple comparisons.



**Fig. 3.** Tract-based registration improves colocalization.

Fig. 3 – Registration of single tracts to the average tract substantially improves the volume of colocalization for groups and for repeated scans in single subjects. Degree of colocalization is a function of threshold (A). All tested regions were similar to this representative example from the Left Visual Cortex: by increasing the stringency of thresholds (eliminating progressively more low-value voxels), the degree of volumetric mismatch between single tracts and the average of all tracts is revealed. The largest improvements in volumetric overlap, both for automated and manual colocalization, occur in the highest-value voxels. Across all cortical regions tested (B), manual colocalization doubles the degree of volumetric overlap. Similarly, following colocalization, individual tracts draw closer to the center of gravity for the average tract, reducing mismatch by 38–66% (C). For subjects scanned twice (D, one month interval), colocalization improved volumetric overlap by a degree similar to that seen in A and B. Automated colocalization also significantly and substantially improves volumetric overlap and reduces location mismatch (all regions reached statistical significance in B and C, but indicators were omitted for clarity). Automated colocalization followed by manual correction of those subjects with the lowest 10% of overlap improves colocalization to nearly match that of manual segmentation (B–D). Multiple-comparisons corrected significance threshold  $p$ -value =  $8 \times 10^{-3}$ . \* indicates  $p$  between  $8.3 \times 10^{-3}$  and  $8.3 \times 10^{-5}$ . \*\* indicates  $p < 8.3 \times 10^{-5}$ . Abbreviations: DSC – Dice similarity coefficient; Avg. – Average.

### 3.5. Location variability in subcortical vs. cortico-callosal tracts

For both the thalamo-putaminal and pallido-rubral tracts, variability in tract location was substantially less than for cortico-callosal tracts. Following whole-brain registration, mean difference in center of gravity for each of the four subcortical tracts was less than 1.1 mm (range: 0.98–1.09, which is less than the resolution of the acquired images), compared with a mean of 3.2 mm for the six cortical tracts assessed (range: 2.52–3.88 mm; Table 3). For the subcortical tracts, whole-brain registration produced tracts that were colocalized as well or better than even manual colocalization of the cortical tracts studied. The subcortical projection with the largest variability in location (right

pallido-rubral tract,  $1.09 \pm 0.10$  mm) was significantly less-variable than the cortico-callosal projection with the lowest variability in location (right visual cortex,  $2.52 \pm 0.22$  mm;  $p < 1.2 \times 10^{-6}$ , Wilcoxon-Mann-Whitney test). Since whole-brain registration colocalized these subcortical tracts significantly better than it colocalized cortico-callosal tracts, one might surmise that registration into tract space would have less impact on group comparisons for subcortical tracts. Indeed, while the peak value of averaged cortico-callosal tracts increased with colocalization (Fig. 2C, Table 3), colocalization of subcortical tracts had little impact on peak value. Following automated colocalization, the peak value of cortico-callosal tractography increased by an average of 28.3%, while for subcortical tracts the peak increased by only 2.2%.

**Table 2**  
Overlap (Dice similarity coefficient) of highest-amplitude 25% of voxels.

A 37 Controls, single tract vs. average tract						
	Left-PMC	Right-PMC	Left-SMA	Right-SMA	Left-VC	Right-VC
Whole-brain registration	11.9 ± 1.7	17.6 ± 2.6	17.0 ± 2.6	5.9 ± 1.5	8.4 ± 1.6	9.5 ± 1.6
Automated Colocalization	23.6 ± 2.1	38.0 ± 3.3	33.0 ± 3.0	21.5 ± 3.1	19.9 ± 2.0	23.2 ± 1.7
Manual Colocalization	25.8 ± 1.9	38.1 ± 2.6	28.8 ± 3.0	27.1 ± 2.5	21.0 ± 1.6	22.7 ± 1.4
p-value, manual colocalization vs. whole-brain registration, t-test	< 7.0x10 <sup>-6</sup>	< 2.7x10 <sup>-6</sup>	< 4.0x10 <sup>-3</sup>	< 6.0x10 <sup>-7</sup>	< 1.8x10 <sup>-6</sup>	< 6.0x10 <sup>-7</sup>
X Fold improvement in overlap, standard vs. manual colocalization	2.2	2.2	1.7	4.6	2.5	2.4

B 14 Controls, repeated scans (scan 1 vs. scan 2)						
	Left-PMC	Right-PMC	Left-SMA	Right-SMA	Left-VC	Right-VC
Whole-brain registration	8.5 ± 2.5	12.6 ± 3.7	17.0 ± 3.9	12.0 ± 2.9	8.4 ± 3.2	10.5 ± 3.0
Automated colocalization	21.9 ± 6.2	18.8 ± 2.7	24.0 ± 5.0	18.5 ± 3.6	12.9 ± 3.9	14.4 ± 3.5
Manual colocalization	26.1 ± 4.4	27.1 ± 3.2	25.2 ± 2.5	22.6 ± 2.5	18.6 ± 3.2	25.8 ± 4.7
p-value, manual colocalization vs. whole-brain registration, t-test	< 4.2x10 <sup>-4</sup>	< 1.4x10 <sup>-3</sup>	< 0.014	< 0.024	< 7.5x10 <sup>-3</sup>	< 7.8x10 <sup>-3</sup>
X Fold improvement in overlap, standard vs. manual colocalization	3.1	2.1	1.5	1.9	2.2	2.5

Table 2 Legend: Tracts registered through the standard, whole-brain method have poor overlap with the average tract (A) and poor overlap between two scans of the same individual (B). Dice similarity coefficients for the tract core (uppermost 25% of voxels) are provided following whole-brain registration, colocalization by our automated colocalization method, and manual colocalization. Overlap between tracts, a key feature in comparing tract amplitude, increases by 50–210% following colocalization. Significance threshold for comparison of Dice similarity coefficients across methods, corrected for multiple comparisons, is  $p < 8.3 \times 10^{-3}$ .

### 3.6. Relative contributions of individual subject and experimental factors to location mismatch

In the 14 subjects who were scanned twice, we isolated the fraction of location mismatch (difference in center of gravity) that was attributable to experimental factors (following the presumption that tract location within an individual is stable over short time intervals) and used that to estimate the fraction of location mismatch that is attributable to inter-subject normal variability. For these 14 subjects, across all six cortico-callosal tracts, the mean difference in center of gravity between individuals was 3.8 mm, while the mean difference in center of gravity between Scan 1 and Scan 2 was 1.7 mm. The mean inter-subject location variability for all cortico-callosal tracts was 55% (range: 39.5–66.1%). Head motion (absolute image displacement, 37 subjects) had no significant impact on DSC (p-values: 0.19–0.91) or tract center of gravity displacement (p-values: 0.30–0.81). Note that head motion is only one potential source of variability in tract location; the absence of correlation between head motion and tract mismatch does not inform other experimental factors that may increase variability in tract location. As the present study was not optimized to distinguish between these other potential confounds, we refer the reader to Landman et al. (2011) for an excellent discussion of these potential experimental factors.

### 3.7. Quantitative assessment of tractography – voxelwise testing of artificial groups

We assessed the effect of tract location on voxelwise tractographic

comparisons using demographically-matched groups of healthy subjects (16 vs. 16). These groups were then shuffled (by flipping the two members of a pair between groups) to produce three categories of comparison: while maintaining demographic matching, groups were either matched for tract value, overtly mismatched for tract value (>25% difference), or intermediate (9–15% difference in tract value).

Voxelwise, non-parametric permutation testing of tracts (utilizing *randomise*) following the standard, whole-brain registration yielded significant group differences even when the differences in value between these groups were negligible (cluster mass threshold = 2.33, Fig. 4, Balanced Groups). Specifically, artificial groups that were designed to have minimal difference in mean tract value (p-values for mean amplitude independent of location, 0.96–0.98) demonstrated significant differences in cluster mass (cluster mass of 981 and 549 for two independent Balanced cohorts). However, these significant clusters fell at the margins of the mean tract, rather than at its peak. Voxelwise assessment of those same tracts, when manually colocalized, reduced group difference in cluster mass by 40-fold, effectively eliminating voxelwise differences between the groups.

Conversely, groups with overt mismatches in tract amplitude (difference between group means of 25% and 39%, Fig. 4, Mismatched Groups) yielded substantially-larger group differences for manually colocalized tracts than for tracts localized by the standard method, whole-brain registration. These included cluster mass values 4–57 times larger for colocalized tracts than for whole-brain registered tracts. The ability to identify voxelwise differences for groups with smaller differences in mean tract value (difference in mean of 9–15%, Fig. 4,



**Table 3**

–Difference in center of gravity (mm) of highest 25% of voxels.

A Cortico-callosal tracts - 37 Controls, individual subject vs. average tract						
	Left-PMC	Right-PMC	Left-SMA	Right-SMA	Left-VC	Right-VC
Whole-brain registration	3.3 ± 0.3	3.1 ± 0.4	2.8 ± 0.3	3.9 ± 0.4	3.0 ± 0.4	2.5 ± 0.2
Automated colocalization	1.6 ± 0.2	1.2 ± 0.1	2.2 ± 0.2	2.5 ± 0.3	1.7 ± 0.2	1.1 ± 0.1
Manual colocalization	1.4 ± 0.2	1.1 ± 0.1	1.8 ± 0.2	2.4 ± 0.2	1.6 ± 0.2	1.1 ± 0.1
p-value, manual colocalization vs. whole-brain registration, t-test	< 5.2x10 <sup>-5</sup>	< 6.0x10 <sup>-7</sup>	< 0.013	< 0.017	< 3.0x10 <sup>-4</sup>	< 6.0x10 <sup>-7</sup>
X Fold decrease in separation, standard vs. manual colocalization	2.4	2.9	1.6	1.6	1.9	2.4
B Subcortical tracts - 37 Controls						
	Left-Thalamo-Putamina	Right- Thalamo-Putamina	Left-Pallido-Rubral	Right-Pallido-Rubral		
Whole-brain registration	0.991 ± 0.11 mm	1.07 ± 0.12	0.977 ± 0.09	1.09 ± 0.10		
Automated colocalization	0.946 ± 0.11	0.995 ± 0.09	0.910 ± 0.10	0.922 ± 0.10		
p-value, automated colocalization vs. whole-brain registration	< 0.80	< 0.63	< 0.65	< 0.28		
C Cortico-callosal tracts – 14 Controls, repeated scans (scan 1 vs. scan 2)						
	Left-PMC	Right-PMC	Left-SMA	Right-SMA	Left-VC	Right-VC
Whole-brain registration	3.2 ± 0.5	2.2 ± 0.5	2.6 ± 0.44	3.5 ± 0.52	3.1 ± 0.75	1.9 ± 0.34
Automated colocalization	2.0 ± 0.46	1.3 ± 0.27	2.8 ± 0.46	2.4 ± 0.47	2.1 ± 0.41	1.3 ± 0.15
Manual colocalization	2.0 ± 0.31	1.3 ± 0.20	2.4 ± 0.49	2.5 ± 0.40	1.5 ± 0.24	1.2 ± 0.13
p-value, manual colocalization vs. whole-brain registration, t-test	< 0.049	< 0.11	< 0.78	< 0.19	< 0.034	< 0.074
X Fold decrease in separation, standard vs. manual colocalization	1.6	1.7	1.1	1.4	2.0	1.6

Table 3 Legend: The center of gravity at the corpus callosum (a measure that combines voxel location and amplitude - akin to center of mass in the physical world) assessed the degree of separation between individual tracts and the average tract (A, B) or between tracts derived from two scans of the same individual (C). For cortico-callosal tracts, colocalization through automated or manual methods reduced the separation in center of gravity. For subcortical tracts, whose passage is constrained within much smaller white matter bundles, the degree of baseline separation and change with colocalization was substantially smaller than for cortico-callosal tracts. Significance threshold, corrected for multiple comparisons, is  $p < 8.3 \times 10^{-3}$ . Abbreviations: PMC – primary motor cortex; SMA – supplemental motor area; VC – visual cortex; Thal-Put – thalamo-putamina; Pall-RN – pallido-rubral; mm – millimeters.

Intermediate Groups) was similarly improved by colocalization, with cluster mass values increasing from zero to 481 in one case, and from 529 to 1472 in another. In contrast with the findings in the Balanced groups, the increased size and significance of clusters for Mismatched and Intermediate groups coincided with the tract core, rather than lying at the margins of the tract. Our aim in these voxelwise assessments was to quantify the impact of tract mismatch on quantitative comparison between groups, not explicitly to describe meaningful differences between these (artificially constructed) groups. Nevertheless, it may be useful to some readers to note that tract colocalization improved the margin by which group comparisons met statistical significance: following colocalization all three Mismatched groups, and one of three Intermediate groups, retained significance when FWE-corrected for multiple comparisons. In contrast, no group retained FWE-corrected significance following whole-brain registration. Balanced groups were not significantly different (following FWE correction) in either whole-brain registration or colocalization conditions.

### 3.8. Quantitative assessment of tractography – amplitude in averaged tracts

To accurately compare amplitude (streamline counts or streamline density) between repeated-session scans in single subjects, tracts must occupy the same location. In subjects scanned twice, averaging tracts derived from first and second scan sessions (within individual subjects) integrated both amplitude and location information; thus, higher mean tract amplitude in the averaged tract indicates reduced loss of signal due to location mismatch. To be explicit, we used this approach to assess the effect of poor colocalization on amplitude assessment, with no assumption that this approach would be used to compare distinct experimental groups; as there is currently no standard tool for the voxelwise assessment of repeated measures tractography, we utilized this approach instead. Note that our method of amplitude quantification (see Materials and Methods 2.7A) parallels the practice of comparing total number or density of streamlines (Chen et al., 2015; Correia et al., 2008; Donahue

**Table 4**  
The impact of tract colocalization on group amplitude comparisons.

A Streamline Count, Average Tract							
Tract	Whole-Brain Registration	Automated Colocalization	Ratio	Tract	Whole-Brain Registration	Automated Colocalization	Ratio
Cortico-callosal tracts							
Left-PMC	43415	54547	1.26	Right-PMC	63933	85437	1.34
Left-SMA	33917	39344	1.16	Right-SMA	35951	46351	1.29
Left-VC	25467	34827	1.37	Right-VC	35087	45161	1.29
Subcortical tracts							
Left-Thal-Put	13588	13844	1.02	Right-Thal-Put	10526	10265	0.975
Left-Pall-RN	3645	3848	1.06	Right-Pall-RN	5755	5972	1.04
B Streamline Density, Average Tract							
Tract	Whole-Brain Registration	Automated Colocalization	Ratio	Tract	Whole-Brain Registration	Automated Colocalization	Ratio
Cortico-callosal tracts							
Left-PMC	33.2	217	6.6	Right-PMC	30.2	180	5.9
Left-SMA	12.7	68.4	5.4	Right-SMA	7.64	68.3	8.9
Left-VC	7.49	60.1	8.0	Right-VC	11.3	88.2	7.8
Subcortical tracts							
Left-Thal-Put	118	118	1.0	Right-Thal-Put	103	83.5	0.81
Left-Pall-RN	24.3	24.3	1.0	Right-Pall-RN	31.6	31.1	0.99

Table 4 Legend: When tracts are averaged across a group, the amplitude of that average is strongly influenced by the degree to which the tract core (the highest 25% of voxels) of each tract overlaps. 37 control subjects were registered through the standard, whole-brain method, and subsequently colocalized through our automated registration method. Comparing the peak value (maximum streamline count, A) or streamline density (B) of these two group averages (standard registration vs. automated colocalization) allowed us to quantify the degree to which tract mismatch compromised tract amplitude. While tract mismatch reduced the streamline count for cortico-callosal tracts (the uppermost panel) by up to 37%, subcortical tracts were not substantially impacted by tract mismatch. Mismatched tracts may underestimate true group differences in amplitude by up to 29% (the mean difference in peak amplitude across the six cortico-callosal regions). Similarly, streamline density was reduced by up to 8.9-fold (mean: 7.1-fold) in cortico-callosal tracts registered through the standard, whole-brain method. Streamline density for subcortical tracts was slightly decreased by colocalization, a difference between groups substantially smaller than the gains in streamline density for cortico-callosal tracts. Note that peak amplitude from the automated colocalization steps (rather than manual colocalization) are presented here so that cortico-callosal tracts can be compared with subcortical tracts. Abbreviations: PMC – primary motor cortex; SMA – supplemental motor area; VC – visual cortex; Thal-Put – thalamo-putaminal; Pall-RN – pallido-rubral.

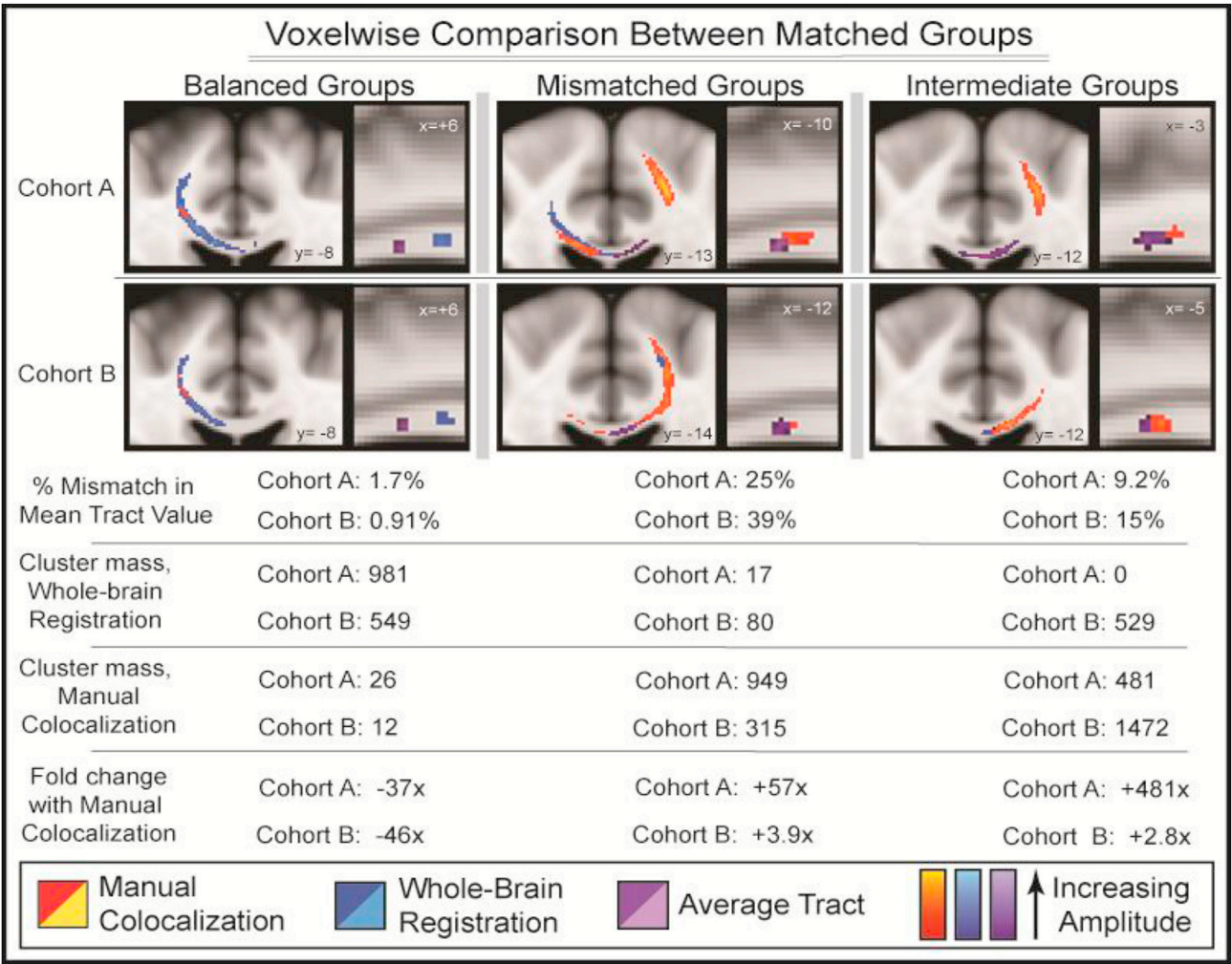
et al., 2016; Hu et al., 2011; Wang et al., 2012). Note also that this approach, which averaged streamline counts across the group, is distinct from the voxelwise comparisons of tracts described above (Results 3.7). Improved tract overlap between repeated scans led to significant increases in average cortico-callosal tract amplitude (Fig. 5A, Table 4). The mean value of all six cortico-callosal tracts increased following colocalization; the value lost to location mismatch reduced by mean 60.6% (range: 53.1–70.0%). P-values were <0.009 for all regions (paired-samples *t*-test, two tailed, whole-brain-registered vs. manually-colocalized tract averages), with 5 of 6 regions maintaining significance when corrected for multiple comparisons (Fig. 5A, p-values of  $9.5 \times 10^{-3}$  -  $1.0 \times 10^{-4}$ ). In addition to improving the accuracy of amplitude assessment at a group level, colocalization increased the accuracy of amplitude measurement for individual subjects (Fig. 5B). For 89% of individual tracts, tract-based registration improved the accuracy of amplitude estimates, while for 6% of tracts the impact on amplitude assessments was neutral, having less than 1% difference between methods. Because some tracts have bifid amplitude distributions, for some individuals and some regions, the tract core straddles the center rather than overlapping it – for those tracts, manual colocalization may lead to a slight decrease in amplitude when assessing the tract core, but an increase in amplitude for lower-threshold tracts (those used in amplitude quantification, e.g. section 3.7). While cortico-callosal tracts uniformly increased in both streamline count and streamline density following colocalization (Table 4), tract-to-tract registration had very little impact on subcortical tracts.

3.9. Quantified diffusion metrics – the impact of tract location

We measured mean FA using thresholded tracts in standard space under two conditions: tracts that were registered through the traditional whole-brain approach, or via manual colocalization. There were no systematic differences in mean FA between these sampling methods: across our six comparisons (two hemispheres, three sites) the mean difference in FA was 1.46%. Four areas had higher FA in colocalized tracts; two areas had higher FA in whole-brain registered tracts. For every area, differences in FA between registration methods were not significant.

4. Discussion

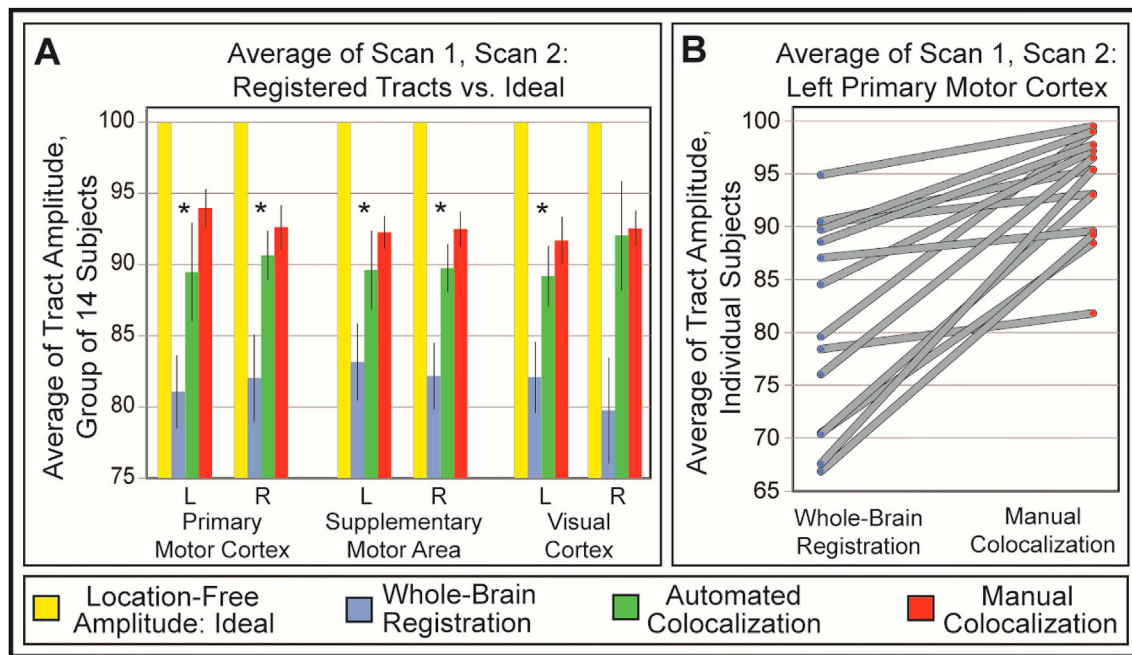
Diffusion tractography is a potent tool for assessing white matter microstructure *in vivo*. However, direct comparison of tractograms derived from injectable tracers and diffusion tractography in mice (Chen et al., 2015) and primates (Azadbakht et al., 2015; Donahue et al., 2016) demonstrated that diffusion tractography is at best a partial reflection of true structural connectivity, and at worst contains false connections and under represents lower-frequency structural connections. With optimization of diffusion tractography parameters, streamlines accurately capture two-thirds to three-quarters of those tracts identified through direct tract tracing (Chen et al., 2015; Donahue et al., 2016), and tractography is >90% accurate in identifying the most-meaningful tracts (Chen et al., 2015). These animal studies remind us that tractographic connectivity does not equal direct measurement of axons (Jones et al., 2013), but also suggest that optimization of tractography for the



**Fig. 4.** Colocalization improves the accuracy of voxelwise testing. Fig. 4 – Voxelwise testing integrates both location and amplitude of tracts; tracts that are misaligned may yield false positive or false negative results from voxelwise testing. We used tractography from the left Primary Motor Cortex to demonstrate the impact that tract location has on voxelwise testing. Pairs of 16 subjects were matched 1-to-1 for age, gender, handedness, and type of scanner. Subject pairs were swapped between groups (maintaining demographic and scanner matching) to produce cohorts (16 vs. 16) for subsequent voxelwise testing in *randomise*. Two examples (Cohort A, Cohort B) are shown for each of three conditions: groups of near-equal amplitude, and groups with large or small differences in amplitude. For groups with near-equal mean tract values (Balanced Groups), whole-brain registration produced false positive results; these false positive results reduced by 40-fold following colocalization (Balanced, Pairs A & B, cluster mass). False positives tended to fall at the margins of the tract distribution, as suggested by their position 4 mm anterior to the average tract (Balanced, sagittal view). In contrast, underestimates of true group differences occurred at the peak of the average tract (Mismatched and Intermediate Groups, sagittal view). For groups differing in mean tract value by 25–39% (Mismatched Groups), voxelwise testing following whole-brain registration underestimated these group differences. Colocalization of tracts improved the ability of voxelwise testing to detect true differences by 4- to 57-fold (Mismatched, Pairs A & B, cluster mass). For groups with small differences in mean tract value (and thus with uncertain likelihood of finding groupwise statistical differences), whole-brain registration eliminated false-negative findings in some cases and improved the detection of true differences in others (Intermediate, Pairs A & B, cluster mass). Images are presented with T-stat range: 2–4. X and Y values indicate the plane of section, MNI coordinates, mm.

investigated structures and imaging metrics holds great promise for investigating structural connectivity in human disease and development. Indeed, streamline counts correlate with clinical measures of disease severity in Parkinson disease (Tan et al., 2015) and autism (Catani et al., 2016), and distinguish Parkinson disease from other causes of tremor (Kim and Park, 2016). Similarly, streamline counts successfully mapped the anatomical infrastructure underpinning the spread of pathology in amyotrophic lateral sclerosis (Schmidt et al., 2016) and suggested a novel excitotoxic mechanism underlying X-linked dystonia parkinsonism (Blood et al., 2017). While whole-brain or large-tract measures of tractography can identify large-scale differences in structural connectivity, seed-based probabilistic tractography is capable of assessing the structural connectivity of smaller regions and lower-frequency white matter connections without requiring that one assume point-by-point certainty

of the route that connection takes. For neurological disorders in which the exact site of pathology is unknown, this ability to dissect sub-projections with an array of defined targets is a key feature. Probabilistic tractography's ability to capture local variations in tract location may make voxelwise and group-average comparisons of tract amplitude difficult, however, since individual tract locations may vary within large white matter bundles (e.g., the corpus callosum) and these location-sensitive assessments rely on tract alignment in order to properly assess amplitude information across a group. We demonstrated that when comparing centers of mass between individual tracts and the average tract, the standard whole-brain registration method produces individual tracts that vary in location from the average tract by a mean of 2.5–3.9 mm, and in some individuals by > 12 mm (Fig. 2D, Table 3). Rather than comparing streamline counts at their maximum amplitude,



**Fig. 5.** Colocalization improves the accuracy of group-average comparisons.

Fig. 5 – Reducing location mismatch by colocalizing tracts improves the accuracy of amplitude measurement in repeated-measures testing at a group level (A) and when assessed in individual subjects (B). 14 subjects were scanned twice, with one month between scans. In A, amplitude (mean streamline counts) was extracted independent of tract location (yellow bars) to yield an ideal amplitude measure; if tracts were perfectly colocalized, amplitude would equal this ideal value. Amplitude assessment in registered tracts (blue, green, or red bars) had the potential to be compromised by location mismatch. Indeed, registration methods that reduced the degree of location mismatch also had amplitude measures closer to the location-free ideal amplitude measure. The standard, whole-brain registration method yielded tracts whose amplitude was underestimated by 16–20%, while manual colocalization, the “gold standard” for localization, reduced this underestimation to 6–8%. For clarity, only significant differences between the whole-brain registration and manual colocalization approaches are shown on the figure. Note, however, that amplitude measures for manually-colocalized tracts (all regions) were significantly reduced relative to the ideal as well – even when tract mismatch is reduced as much as possible, some underestimation of tract amplitude is inevitable in a group average method. Significance threshold, corrected for multiple comparisons:  $p < 0.0083$ . Range of p-values: 0.0095–0.00010. In B, data from individual subjects is presented; these individual data points make up the Left primary motor cortex group from (A). All other regions followed a similar pattern to that in B. Expressed as a percentage of that individual's ideal tract amplitude, each individual's amplitude measure was more accurate when colocalized than when registered through the standard, whole-brain registration method. This suggests that no matter the size of the group, improved tract registration will increase the accuracy of amplitude measurement in tractography. Abbreviations: L = Left; R = Right.

therefore, voxelwise assessment of tracts registered through the whole-brain method compared tracts at a mean of 63–84% below their peak value (Figs. 2C and 3C, Table 1). Mismatches in tract location, and the resulting comparison of the peaks of some tracts with the slopes of others, contributed to false positive results, false negative results, and underestimates of true tract differences (Figs. 4–5, Table 4). These mismatches in tract location have implications beyond voxelwise testing: any of the methods that quantify tractography or diffusion metrics based on proximity to mean tract skeletons or intermittent sampling along a B-spline curve (Smith et al., 2006; O'Donnell et al., 2009; Yendiki et al., 2011) may be compromised by location mismatch.

In groups with minimal differences in tract amplitude (Fig. 4, Balanced Groups), whole-brain registration yielded tracts with significant voxelwise results at the margins of the mean tract, 4–5 mm distant from the tract center. Colocalization reduced these voxelwise results (cluster mass) by 40-fold, arguing that these marginal results were false positives. Likewise, real differences in tract amplitude between groups may be masked by mismatches in tract location (Fig. 4, Mismatched and Intermediate Groups). Manual colocalization of these amplitude-discordant tracts demonstrated significant group differences at the center of the core tract, where differences in peak tract value “should” produce group differences. Colocalization improves the ability of voxelwise testing to identify these true group differences: for groups with a 9–15% difference in mean amplitude, colocalization increased cluster mass results by 3-to-481-fold; for groups mismatched in amplitude by 25–39%, colocalization increased cluster mass results by 4-to-57-fold. Whole brain registration may lead to false negative results in groups

with true differences, and by corollary, colocalization may increase the probability that voxelwise testing will identify true differences in tract amplitude. Tract location is one of several contributors to inter-individual variability in measures of tract amplitude, and other sources of variability may be difficult or impossible to control. We caution readers to be mindful of the fact that colocalization boosted the accuracy of quantification in groups that were matched for age, gender and handedness. We cannot estimate the effectiveness of reducing mismatch when groups are not carefully matched. This recognition is sobering – while tract amplitude had a Gaussian distribution for the group as a whole, flipping subjects between demographically-matched groups readily distorted group amplitude by as much as 43%. Thus, comparison of small groups, or groups that are not demographically matched, via tractography should be undertaken with caution.

The potential for poor colocalization to compromise the accuracy of amplitude measurements affects a second common method for making longitudinal or groupwise tractographic comparisons: analyzing average streamline counts. For subjects scanned twice, whole brain registration led to poor intra-subject tract overlap (mean DSC for tract core, 11.5%; Fig. 3D, Table 2), which improved by 1.5-to-3.1-fold following colocalization (Fig. 3D, Table 2). For any individual subject, tracts registered via the standard, whole-brain technique differed from colocalized tracts in precise location, but were identical in tract amplitude and architecture. Despite this innate similarity, amplitude measures for manually colocalized tracts were significantly closer to the ideal, location-free amplitude measure than tracts aligned by whole-brain registration (Fig. 5A). Improving tract overlap increased the accuracy of amplitude measures



for individual subjects (Fig. 5B) arguing that improved registration has the potential to improve accuracy in experimental groups of all sizes. Location mismatch between scan sessions within an individual is smaller than location mismatch between individuals (1.7 vs. 3.8 mm, respectively), and is presumably smaller than differences between healthy controls and patients with neurological diseases. If this presumption is true, longitudinal and groupwise imaging studies that fail to account for location mismatch may have to overcome a >20% handicap (Fig. 5A) before true amplitude differences can be detected. In contrast, correcting for location mismatch, and thereby reducing variability in amplitude assessments (as evidenced by the uniform increase in mean tract value by colocalization in subjects scanned twice), will increase statistical power, decrease required sample sizes, and thereby reduce the cost and duration of tractographic imaging studies.

What are the origins of variability in tract location? One possibility is that whole-brain registration is simply not accurate. However, visual inspection of all *flirt* and *fniirt*-registered images revealed no gross errors in registration. The fact that subcortical tracts were much more tightly clustered than cortico-callosal tracts, even prior to colocalization, (mean COG difference: subcortical, 1.03 mm; cortico-callosal, 3.2 mm) argues that variance in tract location is not a general feature of whole-brain registration. Likewise, the similar degrees of variance in all cortico-callosal tracts, despite their seed regions being oriented in separate planes and located in distinct areas of the cortex, argues that location mismatch is not simply a consequence of registration failing in particular cortical areas or relative orientations. In short, we find no evidence that whole-brain registration was faulty, and we continue to utilize the FSL registration toolset.

A second potential source of variability in tract location is the size constraints of the transited white matter structure. Spatial heterogeneity in tract location appears to be the norm, as cortico-callosal tracts within our healthy volunteer population differed by as much as 20.1 mm. In large white matter bundles, single axons may have considerable latitude in their precise location within the bundle (Plassard et al., 2015). Indeed, this has been reported previously in the splenium: transcallosal fibers connecting left and right occipital cortices display considerable inter-individual variation in the location (Dougherty et al., 2005). Moreover, tracts connecting discrete cortical sites overlap those connecting other cortical sites in the splenium, both functionally-distinct parts of occipital cortex (Dougherty et al., 2005) and also tracts from occipital, temporal, and parietal lobes (Berlucchi, 2014).

Though tractography follows the composite diffusion properties of axon bundles, not single axons, spatial heterogeneity among the single axons in a bundle will increase the probability of a tracing “jumping” from one fiber to a neighbor. Supporting this premise, we found that variability in tract location was most likely to occur in the long axis (antero-posterior) where the corpus callosum is relatively narrow (primary motor and supplementary motor area tracts), but in the splenium, where the corpus callosum broadens, variability was essentially equal in antero-posterior and superior-inferior orientations (visual cortex tracts). Similarly, tracts passing through much narrower white matter bundles (thalamo-putaminal and pallido-rubral tracts) had variability in tract location that was reduced by 66.9% relative to cortico-callosal tracts. Location mismatch in the best-matched cortical area (following manual colocalization) was essentially equal to the worst-matched subcortical area (following whole-brain registration, 1.06 vs. 1.09 mm, respectively). We hypothesize that increased variability in tract location is a normal feature of tracts passing through large white matter bundles, and is greater in structures that possess greater space for shifts to take place. An alternative hypothesis is that phylogenetically older (e.g., subcortical) tracts have undergone greater evolutionary refinement than neocortical tracts and are therefore more tightly bundled, independent of the size of the transited white matter structure. Note that these two hypotheses are not mutually exclusive, and will require further study to disentangle.

A third potential source of variability in tract location is individual differences in tract architecture (e.g., single-peak vs. divided-peak tracts,

or the degree to which the tract was tilted relative to the sagittal plane). It is notable that certain seed regions were more likely than others to produce bundles with bifid or tripartite peaks: relative to the primary motor and primary visual cortical seeds, the supplementary motor area was approximately twice as likely to produce linearly-arrayed, discrete bundles instead of a single bundle. We find the premise that streamlines might “jump” between bundles an insufficient explanation for these divided tracts for two reasons. First, if such jumping were a general property of tractography we predict that streamlines from all three of our cortical targets would display this property with similar frequency. As noted above, the SMA is much more likely to produce such divided tracts. Second, if streamlines occasionally slipped between bundles those bundles would have to be closely apposed. Instead, we identified individuals whose bifid tracts were separated by multiple millimeters, and for tripartite tracts the peaks may be distributed over a centimeter or more.

Failure to recognize these region-specific sources of variability in tract location could confound whole-brain registration methods, even those that attempt to directly register streamlines at a whole-brain level (Garyfallidis et al., 2015). Indeed, these individual differences in tract architecture were an important confound of our own tract-to-tract registration method. Our automated colocalization method had a tendency to identify one limb of a divided (bifid or tripartite) peak and register that portion as if it were the whole, especially in subjects with asymmetrical bifid peaks (high peak – valley – low peak). These were readily identified in our quality control step (manual colocalization), but would not be obvious from simple comparisons of tract center of gravity or Dice similarity coefficient. A deeper question, not addressed in this manuscript, is whether divided bundles represent structural divisions with important functional distinctions. Individual differences in tract architecture, and the increased variance in tract architecture attributable to particular cortical regions, underscore the importance of visual inspection of one's tractography results. Such quality control measures make a substantial impact: as noted above, improvement of tract registration by even 1–2 mm is associated with substantially more accurate amplitude and volume assessments.

Based on comparison of inter-subject versus intra-subject data, we found that inter-subject variability explained 55% of the mismatch in tract location, with stochastic experimental factors accounting for the remaining 45%. The high degree of inter-subject variability in tract location from our cohort paralleled the findings of Sadeghi et al. (2015), who reported that among large white matter tracts, FA and MD had higher variability in the body and splenium of the corpus callosum, the regions transited by our tracts of interest. It should be noted that measurements of traditional diffusion metrics (e.g., FA, MD) may be negatively affected by location mismatch, but the degree of inaccuracy will be inversely correlated with the size of the mask used for extraction. For example, a 2 mm displacement of a small ROI (<50 voxels) leads to sampling of different tissue for two-thirds of the ROI volume. In contrast, a 2 mm displacement of the tract-based masks utilized in this study samples different tissue in only one-fifth of voxels. Extracting FA or MD values using the tract volume will largely sample the same tissue whether or not the tract is colocalized. Of the six tracts we described, none had FA values that were significantly different between whole-brain registration and colocalization (Results – 3.9). The recent swell of interest in individual differences in neuroimaging has yielded numerous examples in which an individual's strength of functional or structural connectivity correlates with cognitive or behavioral traits. These include studies of individual's resistance to distraction (Yamamoto et al., 2015), ease of learning grammatical rules (Floel et al., 2009), strength of self-esteem (Chavez and Heatherton, 2015), the presence of psychopathic traits among healthy volunteers (Yoder et al., 2015), and even the differential effects of genetic and cultural traits on attention (Pornpattananangkul et al., 2016).

Efforts at large-scale tractographic characterization of major and minor tracts have demonstrated that while some tracts (large and small) are bundled tightly, for other (especially smaller) tracts the heterogeneity

in tract location between individuals precludes accurate quantification (Zhang et al., 2010). It is unknown whether these individual differences in tract location have functional implications. Our work demonstrates, however, that regardless of the direct impact of location on function, spatial heterogeneity can markedly reduce the accuracy of tract quantification and thus limit the accuracy of correlations between structural and behavioral measures in individuals.

It is reasonable to question whether adequate sample size will solve the problem of location mismatch, making colocalization unnecessary. Complicating this assumption is the fact that although tract amplitude and location were independent variables (Pearson's  $r = 0.04$ ), the product of amplitude and location had a large effect on group comparisons: whether the highest-value tract in a group falls at the center or at the margin matters a great deal. We demonstrated that groups of healthy controls differed substantially in the degree of tract location mismatch. The range of variability in tract location between different types of subject – defined by disease status, age, gender, ethnicity or other demographic variable – is unknown. Rather than attempting to model location variance into power calculations, and attempting to balance location variance between disparate groups, reducing location mismatch through colocalization offers the opportunity to minimize that source of variability. Automated colocalization can be completed in approximately 12 min per subject, with an additional 3 min for subjects that require refinement with registration by weighting volumes. With the additional effort to manually register the 10% of subjects having the greatest residual location mismatch, one can achieve colocalization approaching that of the gold standard, manual colocalization of the whole group. The gains in accuracy and statistical power are well worth this additional effort, and promise to improve the quantitative assessment of probabilistic tractography.

One might reasonably question whether the potential for tract mismatch to confound amplitude measures should lead neuroimaging researchers to abandon location-sensitive measures and simply extract amplitude from individual subjects (across regions of interest or from peak voxels) prior to group comparisons. However, such a location-free approach has at least three substantial limitations. First, identifying the site along a tract that produces a group difference in tractography is critical to understanding the clinical impact of that change, and may suggest pathophysiologic mechanisms underpinning the change. Second, spatially-limited but substantial changes in tractography may be diluted to obscurity when values are extracted across all voxels in a tractographic volume. Third, research methods that depend on anatomical group templates to accurately parcellate the anatomic structures of individual subjects, such as quantitative electroencephalography, functional network analysis, and diffusion tractography, are improved by incorporating individual differences in macrostructure and functional specialization into registration and quantification methods (Chechlacz et al., 2015; Hyde et al., 2012; Langs et al., 2016). Including tract location in assessments of probabilistic tractography is of sufficient value that accounting for individual differences in location may help make this valuable technique even more useful.

## 5. Conclusions

By colocalizing individual probabilistic tractograms to a common “tract space” (Fig. 1), we reduced mismatches in tract location (Fig. 3), which improved the ability of both voxelwise (Fig. 4) and group-average (Fig. 5) testing to eliminate false positive results and to identify true differences in mean amplitude. Our registration method reduced variability in tract location by half, but did not completely eliminate mismatch – on average, colocalized cortico-callosal tracts remained 1.5 mm from the mean center of gravity. It is not clear what degree of tract mismatch is tolerable before quantitative comparisons of tractography are compromised. However, if one wishes to compare tract amplitude or volume between groups or across time (including studies of disease treatment or progression), our data argue that every degree of

improvement in colocalization will improve the accuracy of quantification. At a minimum, quantitative tractographic comparisons must recognize what fraction of their results are attributable to true differences in tract amplitude and which are due to differences in tract location.

## Author contributions

JL Waugh: Study concept/design, collected data, analysis and interpretation of data, authored original draft, generated figures.

JK Kuster: Collected data, analysis and interpretation of data.

ML Makhlof: Collected data, analysis and interpretation of data.

JM Levenstein: Collected data, analysis and interpretation of data.

TJ Multhaupt-Buell: Collected data, analysis and interpretation of data.

SK Warfield: Study concept/design, statistical assessment, revised the manuscript for content.

N Sharma: Revised the manuscript for content.

AJ Blood: Study concept/design, collected data, analysis and interpretation of data, revised the manuscript for content.

## Author disclosures

All authors have confirmed that they have no disclosures, no financial relationships with corporations or entities that could potentially benefit from this work, and no conflicts of interest.

## Acknowledgements

Shared computer resources and storage were made possible by the Shared Instrumentation Grants 1S10RR023043 and 1S10RR023401.

Dr. Waugh was supported by the Clinical Research Training Fellowship, American Academy of Neurology, and the Silverman Family Fellowship, Bachmann-Strauss Dystonia and Parkinson Foundation.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.neuroimage.2018.12.057>.

## References

- Abe, O., et al., 2004. Topography of the human corpus callosum using diffusion tensor tractography. *J. Comput. Assist. Tomogr.* 28, 533–539.
- Alexander, A.L., et al., 2001. Analysis of partial volume effects in diffusion-tensor MRI. *Magn. Reson. Med.* 45, 770–780.
- Armitage, P., Berry, G., 1987. *Statistical Methods in Medical Research*, Second Edition, second ed. Blackwell Scientific Publications Ltd.
- Azadbakht, H., et al., 2015. Validation of high-resolution tractography against in vivo tracing in the macaque visual cortex. *Cerebr. Cortex* 25, 4299–4309.
- Behrens, T.E., et al., 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.* 50, 1077–1088.
- Behrens, T.E., et al., 2007. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* 34, 144–155.
- Behrens, T., S., Webster, M., Nichols, T., 2014. *Randomise*. FMRIB. University of Oxford, UK.
- Berlucchi, G., 2014. Visual interhemispheric communication and callosal connections of the occipital lobes. *Cortex* 56, 1–13.
- Blood, A.J., et al., 2006. White matter abnormalities in dystonia normalize after botulinum toxin treatment. *Neuroreport* 17, 1251–1255.
- Blood, A.J., et al., 2012. Evidence for altered basal ganglia-brainstem connections in cervical dystonia. *PLoS One* 7, e31654.
- Blood, A.J., et al., 2017. Increased Insula-Putamen Connectivity in X-Linked Dystonia-Parkinsonism. *NeuroImage: Clinical* (in press).
- Bullmore, E.T., et al., 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imag.* 18, 32–42.
- Caan, M.W., et al., 2011. Nonrigid point set matching of white matter tracts for diffusion tensor image analysis. *IEEE Trans. Biomed. Eng.* 58, 2431–2440.
- Catani, M., et al., 2016. Frontal networks in adults with autism spectrum disorder. *Brain* 139, 616–630.
- Chavez, R.S., Heatherton, T.F., 2015. Multimodal frontostriatal connectivity underlies individual differences in self-esteem. *Soc. Cognit. Affect Neurosci.* 10, 364–370.

- Chechlac, M., et al., 2015. Structural variability within frontoparietal networks and individual differences in attentional functions: an approach using the theory of visual attention. *J. Neurosci.* 35, 10647–10658.
- Chen, H., et al., 2015. Optimization of large-scale mouse brain connectome via joint evaluation of DTI and neuron tracing data. *Neuroimage* 115, 202–213.
- Chouinard, P.A., Paus, T., 2006. The primary motor and premotor areas of the human cerebral cortex. *Neuroscientist* 12, 143–152.
- Ciccarelli, O., et al., 2006. Probabilistic diffusion tractography: a potential tool to assess the rate of disease progression in amyotrophic lateral sclerosis. *Brain* 129, 1859–1871.
- Colby, J.B., Soderberg, L., Lebel, C., Dinov, I.D., Thompson, P.M., Sowell, E.R., 2012. Along-tract statistics allow for enhanced tractography analysis. *Neuroimage* 59, 3227–3242.
- Corouge, I., Fletcher, P.T., Joshi, S., Gouttard, S., Gerig, G., 2006. Fiber tract-oriented statistics for quantitative diffusion tensor MRI analysis. *Med. Image Anal.* 10, 786–798.
- Correia, S., et al., 2008. Quantitative tractography metrics of white matter integrity in diffusion-tensor MRI. *Neuroimage* 42, 568–581.
- Donahue, C.J., et al., 2016. Using diffusion tractography to predict cortical connection strength and distance: a quantitative comparison with tracers in the monkey. *J. Neurosci.* 36, 6758–6770.
- Dougherty, R.F., Ben-Shachar, M., Deutsch, G., Potanina, P., Bammer, R., Wandell, B.A., 2005. Occipital-callosal pathways in children: validation and atlas development. *Ann. N. Y. Acad. Sci.* 1064, 98–112.
- Floel, A., et al., 2009. White matter integrity in the vicinity of Broca's area predicts grammar learning success. *Neuroimage* 47, 1974–1981.
- Gallay, M.N., et al., 2008. Human pallidothalamic and cerebellothalamic tracts: anatomical basis for functional stereotactic neurosurgery. *Brain Struct. Funct.* 212, 443–463.
- Gallichan, D., et al., 2010. Reducing distortions in diffusion-weighted echo planar imaging with a dual-echo blip-reversed sequence. *Magn. Reson. Med.* 64, 382–390.
- Garyfallidis, E., et al., 2015. Robust and efficient linear registration of white-matter fascicles in the space of streamlines. *Neuroimage* 117, 124–140.
- Guye, M., et al., 2003. Combined functional MRI and tractography to demonstrate the connectivity of the human primary motor cortex in vivo. *Neuroimage* 19, 1349–1360.
- Ha, R., Ha, J., 2011. *Integrative Statistics for the Social and Behavioral Sciences*. SAGE Publications, Inc.
- Hayasaka, S., Nichols, T.E., 2004. Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage* 23, 54–63.
- Holodny, A.I., et al., 2005. Diffusion-tensor MR tractography of somatotopic organization of corticospinal tracts in the internal capsule: initial anatomic results in contradistinction to prior reports. *Radiology* 234, 649–653.
- Hu, B., et al., 2011. Quantitative diffusion tensor deterministic and probabilistic fiber tractography in relapsing-remitting multiple sclerosis. *Eur. J. Radiol.* 79, 101–107.
- Hyde, D.E., et al., 2012. Anisotropic partial volume CSF modeling for EEG source localization. *Neuroimage* 62, 2161–2170.
- Innocenti, G.M., et al., 1977. Exuberant projection into the corpus callosum from the visual cortex of newborn cats. *Neurosci. Lett.* 4, 237–242.
- Jenkinson, M., 2014. *FLIRT User Guide. Analysis Group. FMRIB, Oxford, UK*. <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT/UserGuide>.
- Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841.
- Jones, D.K., Cercignani, M., 2010. Twenty-five pitfalls in the analysis of diffusion MRI data. *NMR Biomed.* 23, 803–820.
- Jones, D.K., et al., 2013. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. *Neuroimage* 73, 239–254.
- Kim, M., Park, H., 2016. Using tractography to distinguish SWEDD from Parkinson's disease patients based on connectivity. *Parkinsons Dis* 2016, 8704910.
- Landman, B.A., et al., 2011. Multi-parametric neuroimaging reproducibility: a 3-T resource study. *Neuroimage* 54, 2854–2866.
- Langs, G., et al., 2016 Oct. Identifying shared brain networks in individuals by decoupling functional and anatomical variability. *Cerebr. Cortex* 26 (10), 4004–4014.
- Leemans, A., et al., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. In: 17th Annual Meeting of Intl Soc Mag Reson Med, Hawaii, USA, p. 3537.
- Maddah, M., Grimson, W.E., Warfield, S.K., Wells, W.M., 2008. A unified framework for clustering and quantitative analysis of white matter fiber tracts. *Med. Image Anal.* 12, 191–202.
- Mayer, A., et al., 2011. A supervised framework for the registration and segmentation of white matter fiber tracts. *IEEE Trans. Med. Imag.* 30, 131–145.
- O'Donnell, L.J., Westin, C.F., Golby, A.J., 2009. Tract-based morphometry for white matter group analysis. *Neuroimage* 45, 832–844.
- O'Donnell, L.J., et al., 2012. Unbiased groupwise registration of white matter tractography. *Med Image Comput Comput Assist Interv* 15, 123–130.
- Olivetti, E., et al., 2016. Alignment of tractograms as graph matching. *Front. Neurosci.* 10, 554.
- Park, H.J., et al., 2003. Spatial normalization of diffusion tensor MRI using multiple channels. *Neuroimage* 20, 1995–2009.
- Picard, N., Strick, P.L., 2001. Imaging the premotor areas. *Curr. Opin. Neurobiol.* 11, 663–672.
- Plassard, A.J., et al., 2015. Evaluation of atlas-based white matter segmentation with eve. *Proc. SPIE-Int. Soc. Opt. Eng.* 9413.
- Pornpattananakul, N., et al., 2016. Cultural influences on neural basis of inhibitory control. *Neuroimage* 139, 114–126.
- Putnam, M.C., et al., 2010. Cortical projection topography of the human splenium: hemispheric asymmetry and individual differences. *J. Cognit. Neurosci.* 22, 1662–1669.
- Sadeghi, N., et al., 2015. Analysis of the contribution of experimental bias, experimental noise, and inter-subject biological variability on the assessment of developmental trajectories in diffusion MRI studies of the brain. *Neuroimage* 109, 480–492.
- Schmidt, R., et al., 2016. Simulating disease propagation across white matter connectome reveals anatomical substrate for neuropathology staging in amyotrophic lateral sclerosis. *Neuroimage* 124, 762–769.
- Schwarz, C.G., et al., 2014. Improved DTI registration allows voxel-based analysis that outperforms tract-based spatial statistics. *Neuroimage* 94, 65–78.
- Smith, S.M., et al., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* 31, 1487–1505.
- Swinscow, T., Campbell, M., 1997. *Statistics at Square One*, Ninth Edition, ninth ed. BMJ Publishing Group.
- Talairach, J., Tournoux, P., 1988. *Co-Planar Stereotaxic Atlas of the Human Brain: 3-D Proportional System: an Approach to Cerebral Imaging*. Thieme.
- Tan, W.Q., et al., 2015. Deterministic tractography of the nigrostriatal-nigropallidal pathway in Parkinson's disease. *Sci. Rep.* 5, 17283.
- Tournier, J.D., et al., 2004. Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *Neuroimage* 23, 1176–1185.
- van der Kouwe, A.J., et al., 2005. On-line automatic slice positioning for brain MR imaging. *Neuroimage* 27, 222–230.
- Vos, S.B., et al., 2011. Partial volume effect as a hidden covariate in DTI analyses. *Neuroimage* 55, 1566–1576.
- Wahl, M., et al., 2007. Human motor corpus callosum: topography, somatotopy, and link between microstructure and function. *J. Neurosci.* 27, 12132–12138.
- Wang, J.Y., et al., 2012. A comprehensive reliability assessment of quantitative diffusion tensor tractography. *Neuroimage* 60, 1127–1138.
- White, L.E., et al., 1997. Structure of the human sensorimotor system. I: morphology and cytoarchitecture of the central sulcus. *Cerebr. Cortex* 7, 18–30.
- Winkler, A.M., et al., 2014. Permutation inference for the general linear model. *Neuroimage* 92, 381–397.
- Yamamoto, M., et al., 2015. White matter microstructure between the pre-SMA and the cingulum bundle is related to response conflict in healthy subjects. *Brain Behav* 5, e00375.
- Yeh, F.C., et al., 2013. Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PLoS One* 8, e80713.
- Yendiki, A., et al., 2011. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. *Front. Neuroinf.* 5, 23.
- Yendiki, A., et al., 2013. Spurious group differences due to head motion in a diffusion MRI study. *Neuroimage* 88c, 79–90.
- Yoder, K.J., et al., 2015. Amygdala subnuclei connectivity in response to violence reveals unique influences of individual differences in psychopathic traits in a nonforensic sample. *Hum. Brain Mapp.* 36, 1417–1428.
- Zhang, Y., et al., 2010. Atlas-guided tract reconstruction for automated and comprehensive examination of the white matter anatomy. *Neuroimage* 52, 1289–1301.
- Zollei, L., et al., 2010. Improved tractography alignment using combined volumetric and surface registration. *Neuroimage* 51, 206–213.